УДК 004.89:658.818 DOI https://doi.org/10.35546/kntu2078-4481.2026.2.29

О. А. ЄРОШЕНКО
доктор філософії,
доцент кафедри електронних обчислювальних машин
Харківський національний університет радіоелектроніки
ORCID: 0000-0001-6221-7158

В. М. ФЕДОРЧЕНКО
кандидат технічних наук, доцент,
доцент кафедри електронних обчислювальних машин
Харківський національний університет радіоелектроніки,
доцент кафедри інформаційних систем
Харківський національний економічний університет імені Семена Кузнеця
ORCID: 0000-0001-7359-1460

С. О. ПАРТИКА
старший викладач кафедри електронних обчислювальних машин
Харківський національний університет радіоелектроніки
ORCID: 0000-0002-7376-8980

ОПТИМІЗАЦІЯ ТА ПЕРСПЕКТИВИ РОЗВИТКУ ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ОБСЛУГОВУВАННЯ КЛІЄНТІВ НА БАЗІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

У статті здійснено комплексний аналіз проблематики та перспектив оптимізації інтелектуальних систем обслуговування клієнтів (Intelligent Customer Service Systems) в умовах стрімкої цифровізації бізнес-процесів. Розглянуто сучасні виклики, що постають перед діалоговими системами на прикладі платформи Alibaba Xiaomi, зокрема обмеження у семантичній інтерпретації контексту, обробці неструктурованих запитів «довгого хвоста» та підтримці природності комунікації. Проаналізовано вплив цих факторів на ключові показники ефективності, такі як час відгуку, точність відповідей та рівень задоволеності користувачів.

У роботі запропоновано та обґрунтовано трирівневу стратегію вдосконалення архітектури інтелектуального сервісу з використанням передових можливостей великих мовних моделей. Перший рівень передбачає оптимізацію механізмів управління контекстом на базі архітектури Transformer. Другий рівень фокусується на реалізації повноцінної мультимодальної взаємодії. Описано підходи до інтеграції гетерогенних потоків даних (голос, зображення, текст) у єдиний векторний простір за допомогою моделей на кшталт CLIP, що забезпечує синергетичну обробку візуальних та вербальних запитів у реальному часі. Третій рівень модернізації присвячено розробці адаптивного модуля сентимент-аналізу. Розглянуто застосування алгоритмів глибокого навчання (зокрема, донавчання моделей RoBERTa) для розпізнавання складних емоційних конструктів, таких як сарказм, іронія чи приховане невдоволення. Запропоновано механізм динамічного коригування тональності відповідей та стратегій розв'язання проблем залежно від емоційного стану клієнта.

Окрему увагу в статті приділено майбутнім трендам розвитку галузі, зокрема глибокій інтеграції генеративних моделей із графами знань (Knowledge Graphs). Обґрунтовано, що таке поєднання дозволить підвищити фактологічну точність відповідей, забезпечити можливість логічного виведення та мінімізувати ризики «галюцинацій» нейромереж у критично важливих сферах, таких як медицина, фінанси та освіта. На основі проведеного моделювання та аналізу прогнозних даних (2022–2025 рр.) доведено, що імплементація запропонованих рішень здатна підвищити точність контекстуального розуміння та загальний рівень задоволеності користувачів на 2–6%, що підтверджує практичну цінність дослідження для подальшого розвитку екосистем цифрового обслуговування.

Ключові слова: *інтелектуальні системи обслуговування, великі мовні моделі, обробка природної мови, мультимодальна взаємодія, сентимент-аналіз, графи знань, архітектура Transformer, контекстуальний аналіз.*

O. A. YEROSHENKO
Doctor of Philosophy,
Associate Professor at the Department of Electronic Computers
Kharkiv National University of Radio Electronics
ORCID: 0000-0001-6221-7158

ISSN 2078-4481

V. M. FEDORCHENKO
Doctor of Philosophy, Associate Professor,
Associate Professor at the Department of Electronic Computers
Kharkiv National University of Radio Electronics,
Associate Professor at the Department of Information Systems
Simon Kuznets Kharkiv National University of Economics
ORCID: 0000-0001-6221-7158

S. O. PARTYKA
Senior Lecturer at the Department of Electronic Computers
Kharkiv National University of Radio Electronics
ORCID: 0000-0001-6221-7158

OPTIMIZATION AND FUTURE TRENDS OF LLM-BASED INTELLIGENT CUSTOMER SERVICE SYSTEMS

The article provides a comprehensive analysis of the problems and prospects for optimizing Intelligent Customer Service Systems in the context of rapid business process digitalization. It examines the current challenges facing dialogue systems using the Alibaba Xiaomi platform as an example, particularly limitations in semantic context interpretation, processing unstructured "long-tail" queries, and maintaining the naturalness of communication. The impact of these factors on key performance indicators, such as response time, response accuracy, and user satisfaction level, is analyzed.

The paper proposes and substantiates a three-level strategy for improving the architecture of intelligent customer service using the advanced capabilities of Large Language Models (LLMs). The first level involves optimizing context management mechanisms based on the Transformer architecture. The second level focuses on implementing fully-fledged multimodal interaction. It describes approaches to integrating heterogeneous data streams (voice, image, text) into a single vector space using models like CLIP, which ensures the synergistic processing of visual and verbal queries in real time. The third level of modernization is dedicated to developing an adaptive sentiment analysis module. It considers the application of deep learning algorithms (in particular, the fine-tuning of RoBERTa models) to recognize complex emotional constructs such as sarcasm, irony, or hidden dissatisfaction. A mechanism for dynamically adjusting the tone of responses and problem-solving strategies depending on the client's emotional state is proposed.

Special attention in the article is paid to future industry development trends, in particular, the deep integration of generative models with Knowledge Graphs. It is substantiated that such a combination will improve the factual accuracy of responses, provide the capability for logical inference, and minimize the risks of neural network "hallucinations" in critical domains such as medicine, finance, and education. Based on the conducted modeling and analysis of predictive data (2022–2025), it is proven that the implementation of the proposed solutions can increase the accuracy of contextual understanding and the overall level of user satisfaction by 2-6%, which confirms the practical value of the study for the further development of digital service ecosystems.

Key words: *intelligent customer service systems, large language models, natural language processing, multimodal interaction, sentiment analysis, knowledge graphs, Transformer architecture, contextual analysis.*

Постановка проблеми

Протягом останніх років, на тлі стрімкої еволюції технологій штучного інтелекту, інтелектуальні системи обслуговування клієнтів (ІСОК) трансформувалися у критично важливий елемент цифрової інфраструктури сучасних підприємств. Завдяки інтеграції таких інструментів, як розпізнавання мовлення, обробка природної мови (NLP) та машинне навчання, ці системи забезпечують користувачам миттєвий та високопродуктивний сервіс. У порівнянні з традиційними моделями підтримки, ІСОК демонструють суттєві переваги в аспектах операційної ефективності, оптимізації витрат та покращення користувацького досвіду [1].

Потужного імпульсу розвитку галузі надала поява великих мовних моделей (LLM), таких як GPT від OpenAI та BERT від Google. Навчені на колосальних масивах даних, ці моделі здатні інтерпретувати складний контекст, генерувати діалоги, максимально наближені до природних, та надавати експертні відповіді у різноманітних предметних областях [2]. У зв'язку з цим, питання ефективної імплементації LLM в архітектуру систем обслуговування для підвищення якості реакції та задоволеності клієнтів набуває статусу актуальної науково-практичної проблеми.

Еталонним прикладом реалізації є система «Alibaba Xiaomi», розроблена лідером китайського технологічного ринку. Вона знайшла широке застосування в електронній комерції, охоплюючи сценарії від консультацій щодо товарів до післяпродажного супроводу [2]. Поєднуючи NLP, великі дані та графи знань, «Alibaba Xiaomi» демонструє потенціал інтелектуальних систем на практиці. Втім, зростання вимог до персоналізації та ускладнення запитів користувачів свідчать про те, що наявні рішення потребують подальшого вдосконалення.

Відповідно до функціоналу та методів технічної реалізації, інтелектуальні системи обслуговування можна класифікувати за кількома критеріями. За способом взаємодії виділяють текстові системи, які є найпоширенішою

формою і передбачають отримання відповідей через введення запитань [3]. Голосові системи використовують технології розпізнавання та генерації мови для ведення розмов (наприклад, Siri від Apple, Alexa від Amazon) і підходять для розумних будинків та автомобільних сервісів, звільняючи руки користувачів. Мультимодальні системи поєднують текст, голос, зображення та інші технології для забезпечення різноманітних методів взаємодії.

Швидкий розвиток інтелектуальних систем обслуговування в останні роки дозволив багатьом галузям здійснити цифрову трансформацію клієнтського сервісу. На даний момент основні системи на ринку представлені зрілими продуктами відомих вітчизняних та зарубіжних компаній, кожна з яких має свої особливості в технічній архітектурі та функціоналі. На міжнародному ринку можна виділити чотири основні системи. Серія ChatGPT від OpenAI базується на моделі GPT, а її остання версія GPT-5 демонструє чудові можливості розуміння та генерації природної мови, справляючись із багатоходовими діалогами та емоційною комунікацією [1]. Google Dialogflow – це платформа розробки на базі NLU, що поєднує розпізнавання та синтез мовлення для мультимодальної взаємодії та широко використовується в електронній комерції та фінансах. Azure Bot Service від Microsoft – це хмарний сервіс, що спирається на когнітивні сервіси для побудови корпоративних ботів з підтримкою багатомовності та складних бізнес-процесів. Watson Assistant від IBM вирізняється потужним семантичним розумінням та акцентом на безпеці даних, що робить його придатним для банківської та державної сфер [2].

На китайському ринку присутні щонайменше п'ять провідних систем. «Ali Little Honey» (Alibaba Xiaomi) широко використовується на платформах Taobao та Tmall, поєднуючи NLP та графи знань для точного розуміння намірів користувачів та обробки понад 90% запитів. Система Jingxiaozhi від JD.com охоплює логістику та післяпродажне обслуговування, вміючи обробляти мільйони запитів у пікові періоди. Tencent Cloud Intelligent Customer Service базується на хмарній платформі та підтримує взаємодію через WeChat і QQ, приділяючи увагу аналізу емоцій. Інтелектуальна система Meituan обслуговує доставку їжі та туризм, використовуючи мультимодальну взаємодію для вирішення спорів. Система Didi Chuxing фокусується на транспортних послугах, інтегруючи GPS та голосові технології для швидкого реагування на потреби водіїв та пасажирів [4].

Аналіз ринку показує, що розвиток інтелектуальних систем рухається в напрямку використання великих мовних моделей, впровадження мультимодальної взаємодії, посилення уваги до аналізу емоцій та персоналізації, а також забезпечення безпеки даних через локальне розгортання. Успішна практика згаданих систем, зокрема аналіз Alibaba Xiaomi, надає цінний досвід для подальшої оптимізації сервісів на базі великих мовних моделей.

Аналіз останніх досліджень і публікацій

Динаміка розвитку технологій інтелектуальної підтримки користувачів нерозривно пов'язана з еволюцією їхньої технічної архітектури. Сучасні системи, що функціонують на основі великих мовних моделей, являють собою синтез методів обробки природної мови, алгоритмів глибокого навчання та прикладних бізнес-сценаріїв. Такий комплексний підхід забезпечує наскрізну підтримку процесу обслуговування: від глибинного семантичного розпізнавання запиту до генерації змістовної та контекстуально доречної відповіді [4].

Типова структурна організація інтелектуальних систем обслуговування на базі LLM передбачає наявність п'яти ключових функціональних модулів. Першим ланцюгом є модуль обробки вхідних даних, основна функція якого полягає у прийомі та попередній обробці інформації від користувача, чи то текст, голос, чи зображення. Для голосових повідомлень застосовуються технології перетворення мовлення в текст (STT), тоді як текстові дані підлягають сегментації та стандартизації. Наступним етапом є робота модуля розпізнавання намірів та заповнення слотів (slot filling). Цей компонент виконує кілька критично важливих функцій: за допомогою семантичного аналізу він ідентифікує потреби користувача (наприклад, «повернення товару» або «запит щодо замовлення»), а також вилучає ключові сутності (номер замовлення, назва продукту, час тощо) [5]. Технічна реалізація цього процесу зазвичай базується на використанні класифікаторів, тонко налаштованих (fine-tuned) на моделях типу BERT або GPT, що гарантує високу точність розпізнавання.

За логіку взаємодії відповідає модуль управління діалогом, завданням якого є підтримання контексту протягом кількох ітерацій спілкування та забезпечення чіткої логіки інтеракції між користувачем і системою. Робота цього менеджера базується на алгоритмах відстеження стану діалогу (DST) у поєднанні з можливостями LLM для генерації чутливих до контексту реакцій. Безпосереднє формування відповіді здійснює модуль генерації, який використовує потужності генеративних моделей (наприклад, GPT-5) для створення плавної, семантично зв’язної мови на основі вхідних даних та контексту. Замикає архітектурний цикл модуль зворотного зв'язку та навчання, що дозволяє оптимізувати роботу системи на основі реакцій користувачів. Оцінки якості відповідей, надані клієнтами, використовуються як дані для безперервного донавчання та ітераційного вдосконалення моделі [6].

Технологічна імплементація моделі базується на трьох фундаментальних аспектах. По-перше, це розуміння природної мови (NLU), що реалізується через моделі на базі BERT, здатні аналізувати складні синтаксичні структури та визначати інтенції. По-друге, це генерація природної мови (NLG), де генеративні моделі рівня GPT-5 забезпечують створення високоякісного тексту завдяки попередньому навчанню та подальшому точному налаштуванню. По-третє, важливим елементом є мультимодальна підтримка, яка завдяки використанню архітектури Transformer дозволяє системі опрацьовувати різнорідні вхідні дані, поєднуючи, наприклад, зображення та текст

[4-5]. Як демонструють дані таблиці 1, хоча такі системи, як Alibaba Xiaomi, демонстрували високу ефективність ще в епоху традиційних рішень, інтеграція з великими мовними моделями дозволила їм досягти значного прогресу та суттєвих покращень у всіх аспектах функціонування.

Аналіз порівняльних даних, наведених у таблиці 1, засвідчує суттєву перевагу інтелектуальних систем нового покоління над традиційними рішеннями. Зокрема, інтеграція великих мовних моделей дозволила підвищити точність розпізнавання намірів користувачів на 15 відсоткових пунктів (з 80% до 95%), а показник логічної узгодженості діалогу зріс на 20% (з 70% до 90%), що є критично важливим для підтримки складних комунікаційних сценаріїв. Як наслідок, це призвело до зростання загального рівня задоволеності користувачів до 95% та збільшення частки успішно вирішених запитів до 96%.

На рисунку 1 наочно відображено структуру розподілу ринку інтелектуальних систем обслуговування клієнтів на базі великих мовних моделей у розрізі різних галузей промисловості. Узагальнюючи аналіз функціональних можливостей, можна констатувати, що фундаментальна конкурентна перевага таких систем полягає у їхній здатності до ефективного семантичного розуміння, що дозволяє з високою точністю інтерпретувати та аналізувати складні синтаксичні конструкції. Крім того, критично важливою характеристикою є механізм відстеження контексту, який забезпечує відмінну результативність у підтримці багатоходових діалогів, а також високий рівень масштабованості, що проявляється у підтримці багатомовності та адаптації до різноманітних сценаріїв використання.

Таблиця 1

Порівняльний аналіз показників ефективності системи Alibaba Xiaomi

Показник	Традиційна система, %	Інтелектуальна система обслуговування на базі великих мовних моделей, %
Точність розпізнавання намірів	80	95
Логічна узгодженість (когерентність) багатоетапного діалогу	70	90
Рівень задоволеності користувачів	85	95
Коефіцієнт вирішення проблем	88	96

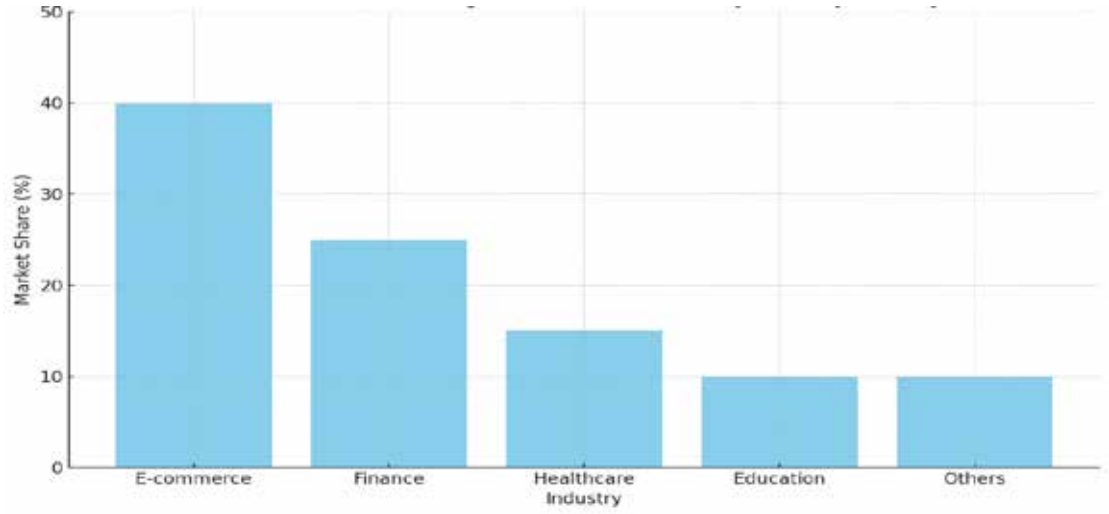

Рис. 1. Частка ринку інтелектуальних систем обслуговування клієнтів за галузями

Водночас, необхідно брати до уваги низку обмежень та викликів, притаманних даній технології. Насамперед, великі мовні моделі характеризуються значною ресурсоємністю, оскільки процеси їх навчання та операційного розгортання вимагають наявності високопродуктивного апаратного забезпечення [7]. Суттєвим аспектом також залишаються питання приватності та інформаційної безпеки, особливо в контексті обробки чутливих персональних даних користувачів, що зумовлює нагальну необхідність впровадження посилених протоколів захисту інформації. Окрім цього, потребує подальшого вдосконалення механізм вирішення рідкісних проблем (так званих проблем «довгого хвоста»), адже, незважаючи на широку базу знань та покриття тем, LLM все ще можуть стикатися з бар'єрами розуміння у нестандартних або специфічних ситуаціях.

Інтелектуальні системи обслуговування клієнтів, будучи одним із ключових напрямів застосування технологій штучного інтелекту, відіграють стратегічну роль у сервісній інфраструктурі сучасних підприємств. Завдяки

конвергенції методів обробки природної мови, глибинного навчання та графів знань, ці системи забезпечують користувачам високопродуктивний та інтелектуалізований досвід взаємодії [8]. Незважаючи на широку імплементацію у різноманітних галузях промисловості, технологія все ще стикається з низкою викликів технічного та прикладного характеру, які потребують вирішення для подальшого масштабування.

Аналіз галузевої специфіки демонструє значну цінність цих систем у низці секторів економіки. У сфері електронної комерції пріоритетними завданнями є консультаційна підтримка щодо продуктів, трекінг логістичних операцій та післяпродажний сервіс. Інтеграція технологій розуміння природної мови дозволяє автоматизувати обробку високочастотних запитів, таких як перевірка статусу доставки, делегуючи складні випадки операторам [9]. У фінансовому секторі, зокрема в банківській справі та страхуванні, системи асистують у процесах кредитування та інвестування [10], використовуючи семантичний аналіз для ідентифікації потреб клієнтів та надання персоналізованих рекомендацій. Медична галузь застосовує інтелектуальних агентів для реєстрації пацієнтів, первинних консультацій та рекомендацій ліків, де критично важливим є швидкий аналіз великих масивів даних. В освітньому секторі платформи використовують історію навчання користувачів для адаптивного підбору курсів, а в туризмі та транспорті системи динамічно коригують пропозиції готелів чи авіарейсів на основі аналізу уподобань клієнтів у реальному часі.

Узагальнення досвіду використання інтелектуальних систем дозволяє виокремити п'ять основних проблемних аспектів їх функціонування. Першим технічним бар'єром є обмеженість семантичного розуміння та утримання контексту у тривалих діалогах, що є "вузьким місцем" існуючих технологій. Цю проблему можна подолати шляхом інтеграції потужних контекстуальних можливостей великих мовних моделей, що забезпечить когерентність багатоходових розмов. Другою проблемою є недостатня ефективність обробки «довгого хвоста» (long tail problem): системи відмінно справляються зі стандартизованими запитами високої частотності, але втрачають точність при роботі з рідкісними, неструктурованими випадками. Вирішенням може стати комбінація механізмів динамічного оновлення баз знань та методів трансферного навчання.

Третій аспект стосується розпізнавання емоцій, що безпосередньо впливає на користувацький досвід [11]. Система повинна не лише механічно відповідати, а й адаптувати стратегію реагування (наприклад, заспокоювати клієнта при скаргах на затримку логістики) за допомогою інтегрованих модулів аналізу тональності. Четвертим викликом є складність мультимодальної взаємодії, оскільки існуючі методи злиття різнорідних даних (голос, текст, зображення) потребують вдосконалення [12-13]. Перспективним є впровадження мультимодальних трансформерів для створення уніфікованої рамки взаємодії [14]. Нарешті, критичного значення набувають питання безпеки даних та приватності, особливо при роботі з конфіденційною фінансовою чи медичною інформацією, що вимагає застосування локального розгортання або технологій федеративного навчання для мінімізації ризиків витоку даних.

Аналіз наведених даних (таблиця 2) свідчить про наявність прямої кореляції між типовістю запитів та ефективністю їх обробки. Найбільшу частку у структурі звернень (60%) займають високочастотні, стандартизовані запитання, для яких система демонструє найвищий рівень успішного закриття – 95%. Водночас, проблеми категорії «довгого хвоста», попри їх незначну питому вагу у загальному потоці (10%), залишаються найскладнішим викликом для автоматизації, оскільки лише половина з них (50%) вирішується системою успішно без залучення оператора.

Таблиця 2

Показники оптимізації інтелектуального обслуговування на основі даних

Тип проблеми	Питома вага (%)	Коефіцієнт успішного вирішення (%)
Високочастотні запити	60	95
Запити середньої частоти	30	85
Проблеми «довгого хвоста» (низькочастотні)	10	50

Крім аналізу успішності вирішення запитів, критичним показником якості сервісу є оперативність. Взаємозв'язок між рівнем задоволеності користувачів та часом відгуку системи наочно продемонстровано на рисунку 2.

Узагальнюючи викладений матеріал, можна стверджувати, що інтелектуальні системи обслуговування підтвердили свою високу прикладну ефективність у стратегічно важливих галузях, таких як електронна комерція, фінансовий сектор, сфера охорони здоров'я та освіта. Водночас, подальша інтеграція цих технологій стикається з низкою суттєвих перешкод, серед яких найбільш критичними є обмеження у глибинному семантичному аналізі, складнощі з обробкою нетипових запитів (проблеми «довгого хвоста»), забезпечення ефективної мультимодальної інтеракції, а також гарантування інформаційної безпеки [15]. Перспективи вирішення окреслених проблем лежать у площині синергії великих мовних моделей із передовими технологічними рішеннями, такими як динамічні бази знань, мультимодальні архітектури трансформерів та методи федеративного навчання. Очікується,

що така інтеграція дозволить суттєво модернізувати сервісний функціонал та оптимізувати користувацький досвід, створюючи надійний технологічний фундамент для процесів цифрової трансформації в різних секторах економіки.

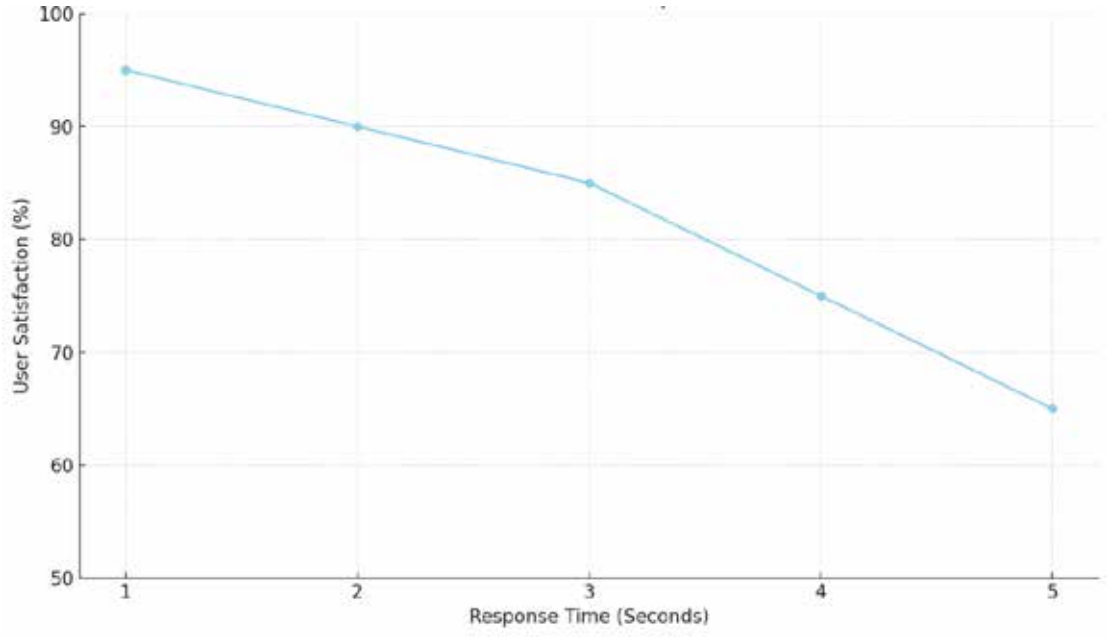

Рис. 2. Залежність рівня задоволеності користувачів від часу реакції системи

Формулювання мети дослідження

Метою статті є розробка теоретичного базису та аналіз практичних аспектів застосування інтелектуальних систем підтримки клієнтів, побудованих на основі великих мовних моделей. Шляхом глибокого аналізу системи «Alibaba Xiaomi» передбачається виявлення сильних та слабких сторін її (систем підтримки клієнтів) технічної реалізації, а також формулювання рекомендацій щодо її вдосконалення.

Для досягнення мети поставлено такі **завдання**:

1. Сформувати теоретичну структуру ICOK на базі LLM, дослідити її ключові технологічні принципи та вектори розвитку.

2. Провести аналіз ефективності «Alibaba Xiaomi» в розрізі технічної архітектури, сценаріїв використання та користувацького досвіду, узагальнивши успішний досвід та напрями оптимізації.

3. Розробити методики та стратегії оптимізації інтелектуальних систем для підвищення рівня задоволеності користувачів та сервісної ефективності, що слугуватиме орієнтиром для інших підприємств.

Викладення основного матеріалу дослідження

Удосконалення механізмів контекстуального розуміння моделі.

У сучасних системах інтелектуального обслуговування здатність моделі коректно інтерпретувати контекст є визначальним фактором, що безпосередньо впливає на логічну зв'язність (когерентність) та точність багатоходових діалогів. Хоча існуюча система Alibaba Xiaomi володіє базовими можливостями відстеження контексту, в умовах складних сценаріїв все ще спостерігаються проблеми з виявленням семантичних зв'язків між репліками, що призводить до логічних розривів та помилок при зміні намірів користувача. Впровадження вдосконалень на основі архітектури Transformer здатне суттєво модернізувати можливості управління контекстом, підвищуючи якість комунікації та рівень задоволеності клієнтів. Аналіз роботи Alibaba Xiaomi показує, що система ефективна в простих діалогах, проте демонструє слабкість при обробці взаємозв'язків між різними етапами розмови. Наприклад, якщо після запитання про обсяг пам'яті конкретного смартфона користувач запитає: «Чи є такий самий білий в наявності?», система може не встановити асоціативний зв'язок між прикметником «білий» та згаданим раніше об'єктом («цей телефон»). Крім того, при багаторазовій зміні теми, наприклад, переході від питання ціни комп'ютера до можливості розстрочки, алгоритм може помилково співвіднести запит про оплату з іншим продуктом.

Для вирішення окреслених проблем доцільно застосувати низку технічних рішень на базі архітектури Transformer. Першим кроком є впровадження механізму динамічної пам'яті, що дозволяє ієрархічно зберігати контекстуальну інформацію: ключові дані – у довгостроковій пам'яті, а поточні – у короткостроковій. У поєднанні

з механізмом уваги (attention mechanism), історичний контент діалогу фільтрується через модуль багатоканальної самоуваги (multi-head self-attention) для збереження сутнісної інформації. Динамічне оновлення стану пам'яті гарантує безперервне відстеження важливих контекстів, що дозволяє системі, наприклад, точно запам'ятовувати атрибути всіх товарів, якими цікавився користувач під час сесії.

Наступним важливим кроком є розширення контекстного вікна, що збільшить фокус моделі на історії діалогу та забезпечить обробку довгих послідовностей. Використання вдосконаленої моделі Transformer із механізмом розрідженої уваги (sparse attention) дозволить знизити обчислювальну складність при роботі з великими обсягами даних. Сегментація історичних діалогів з акцентом на фрагментах, релевантних поточному раунду, суттєво покращить зв'язність розмови та точність відстеження інтенцій. Паралельно необхідно здійснити оптимізацію вагових коефіцієнтів та асоціації намірів. Введення «ваги теми» (Topic Weights) у механізм уваги дозволить пріоритезувати основні наміри користувача. Інтеграція технології відстеження стану діалогу (DST) забезпечить динамічний запис та асоціацію слотів інформації з контекстом. Це дозволить системі швидко перемикати фокус, наприклад, при переході від запиту про замовлення до процесу повернення, зберігаючи прив'язку до конкретної транзакції. Ефективність таких заходів підтверджується даними, наведеними у таблиці 3, де відображено результати оптимізації на основі реальних даних взаємодії.

Таблиця 3

Динаміка точності багатоходових діалогів до та після оптимізації

Показник	До оптимізації (%)	Після оптимізації (%)
Точність визначення контексту	78	92
Точність асоціації намірів користувача	85	96
Логічна узгодженість (когерентність) діалогу	80	94

Для верифікації ефективності моделі рекомендується використовувати стандартні набори даних (наприклад, серії DSTC) та проводити онлайн A/B тестування.

Прогнозовані результати впровадження запропонованих заходів візуалізовано на рисунку 3.

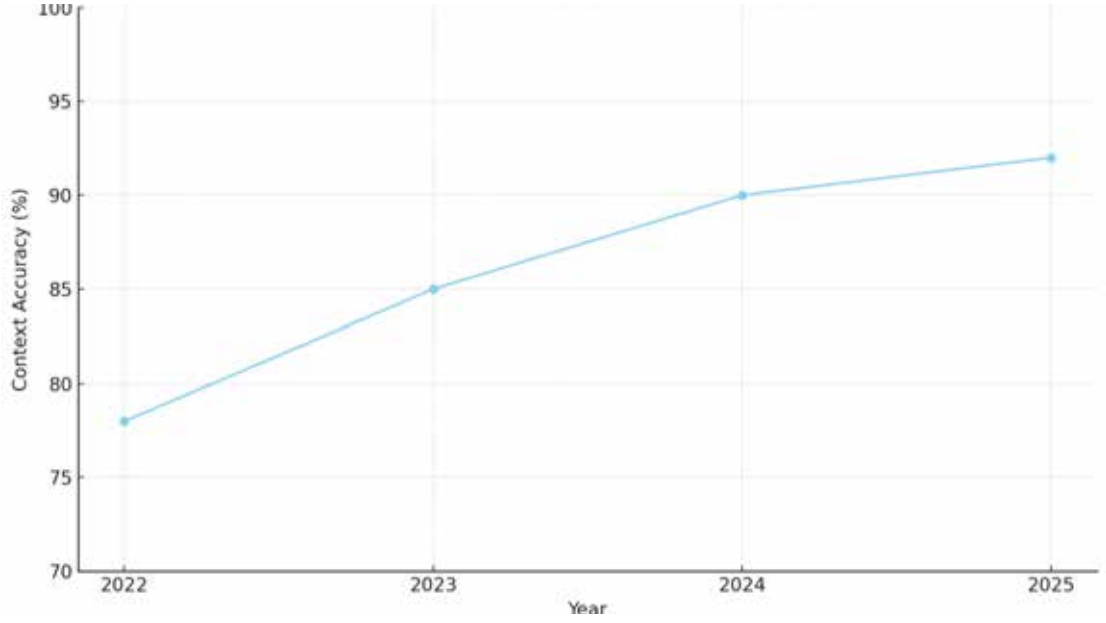

Рис. 3. Прогнозована динаміка зростання точності контекстуального аналізу (2022–2025 рр.)

Очікується, що реалізація вищезазначених стратегій призведе до комплексного позитивного ефекту. По-перше, значно покращиться користувацький досвід завдяки підвищенню плавності діалогу та відчуттю «інтелектуальності» сервісу, здатного миттєво надавати відповіді на пов'язані запитання. По-друге, зросте ефективність обслуговування за рахунок швидшої обробки складних розмов та зменшення необхідності повторних уточнень з боку клієнта. По-третє, вдосконалене управління контекстом сприятиме зростанню показника задоволеності користувачів на 2-5%. Таким чином, оптимізація контексту на базі архітектури Transformer дозволить системі Alibaba Xiaomi точніше фіксувати історію взаємодії, підвищуючи загальну ефективність вирішення проблем.

Імплементація мультимодальної взаємодії

В умовах зростаючої диверсифікації користувацьких запитів, імперативом розвитку інтелектуальних систем обслуговування стає підтримка багатовимірної інтеракції, що охоплює різноманітні модальності вхідних даних: голос, зображення та текст. Однак, існуючий функціонал систем характеризується недостатньою спроможністю до консолідації та комплексної обробки різнорідних даних, що ускладнює ефективну роботу в складних сценаріях. Реалізація повноцінної мультимодальної взаємодії не лише оптимізує користувацький досвід, але й суттєво розширює спектр застосування технології, пропонуючи гнучкі та ефективні рішення для різних галузей.

Аналіз поточної архітектури Alibaba Xiaomi свідчить про домінування текстового каналу комунікації; незважаючи на номінальну підтримку голосу та зображень, можливості їх синергетичного аналізу залишаються обмеженими. Системі бракує уніфікованої мультимодальної моделі, здатної синтезувати інформацію з різних джерел, що призводить до неузгодженості реакцій або часових затримок. Фундаментальна проблема полягає у відсутності кореляції між результатами семантичного аналізу мовлення та візуального розпізнавання об'єктів. Це унеможливлює надання інтегрованих сервісів у таких сферах, як охорона здоров'я, освіта чи електронна комерція, де однієї модальності часто недостатньо для вирішення запиту.

Стратегія вдосконалення мультимодальної взаємодії передбачає реалізацію трьох ключових векторів. Першочерговим є впровадження мультимодальних моделей трансформерів (на кшталт CLIP або Flamingo), які володіють потужним потенціалом крос-модального розуміння. Інтеграція таких архітектур у технічний фреймворк Alibaba Xiaomi дозволить реалізувати злиття інформації через механізм спільної уваги, уніфікуючи візуальні, акустичні та текстові дані в єдиному просторі репрезентацій. Наприклад, при запиті про наявність кольорів для завантаженого зображення товару, система зможе генерувати точну відповідь, базуючись на синтезі візуальних даних та голосового запиту. Другим вектором є оптимізація конвеєра обробки даних шляхом побудови ефективних механізмів попередньої та постобробки, а також створення фреймворку реального часу для синхронізації потоків мовлення, зображень і тексту. Третій напрям стосується посилення емоційного інтелекту системи: інтеграція модулів мультимодального аналізу настроїв (з урахуванням інтонації, виразу обличчя та тексту) на базі моделей глибокого навчання, таких як MMEmo, дозволить системі ідентифікувати незадоволення клієнта та адаптувати сценарій реагування, уникаючи механічних відповідей на користь перемикання на оператора.

Практична імплементація зазначених рішень, враховуючи значні обсяги ринків e-commerce та медицини, здатна принести суттєві вигоди для системи Alibaba Xiaomi. Очікувані кількісні показники ефективності представлено в таблиці 4.

Таблиця 4

Порівняльний аналіз показників ефективності до та після впровадження мультимодальної взаємодії

Показник	До впровадження	Після впровадження
Рівень задоволеності користувачів	88%	95%
Коефіцієнт успішності багатоходових діалогів	85%	93%
Затримка обробки даних (мс)	200	150

Як свідчать наведені дані, інтеграція мультимодальних можливостей сприяє зростанню рівня задоволеності клієнтів на 7 відсоткових пунктів та підвищенню успішності діалогів, при одночасному зменшенні латентності системи на 25%, що наочно ілюструє позитивний вплив запропонованих оптимізаційних заходів. Візуалізація основних ефектів від покращення мультимодальної взаємодії наведена на рисунку 4.

Реалізація окреслених заходів дозволить досягти якісних зрушень за трьома ключовими напрямами: оптимізація користувацького досвіду, підвищення ефективності сервісних операцій та розширення спектра прикладних сценаріїв. У частині взаємодії з користувачем впровадження мультимодальності забезпечить системі необхідну інтуїтивність та гнучкість, що дозволить ефективно задовольняти різнопланові запити клієнтів. Здатність до уніфікованої обробки гетерогенних даних – голосу, зображень та тексту – відкриває шлях до масштабування інтелектуальних систем у таких сферах, як охорона здоров'я, електронна комерція та освіта. Паралельно синхронна обробка мультимодальних потоків сприятиме зростанню операційної ефективності, мінімізуючи час очікування відповіді та прискорюючи процес вирішення проблем.

Для системи Alibaba Xiaomi інтеграція мультимодальних моделей трансформерів, оптимізація конвеєрів даних та вдосконалення модулів розпізнавання емоцій стануть каталізатором проривних змін у складності та варіативності інтеракцій. Ця трансформація не лише комплексно покращить якість обслуговування, але й суттєво розширить функціональні межі системи, надаючи потужну технологічну підтримку для цифрових сервісів підприємства.

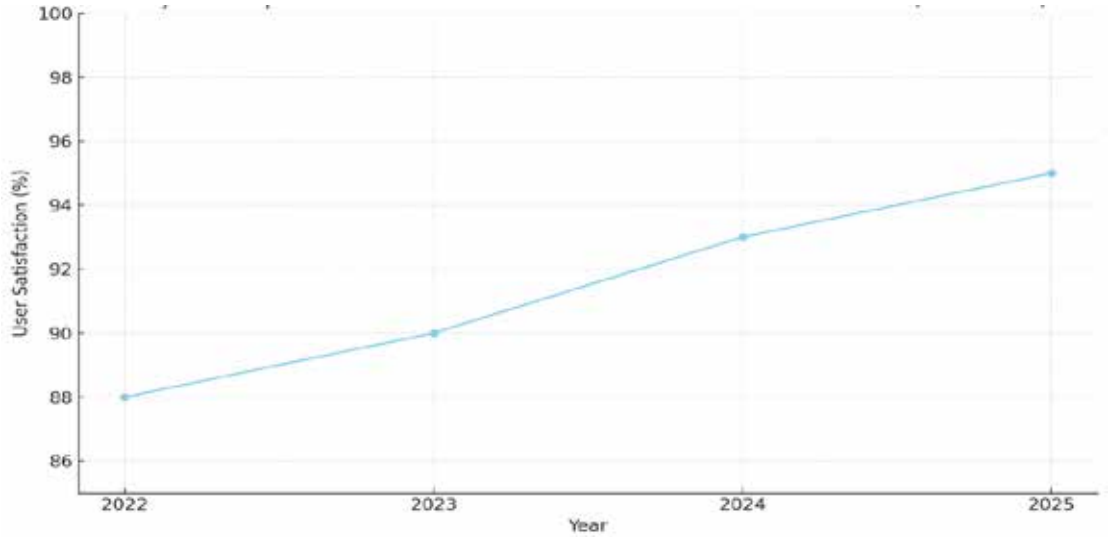

Рис. 4. Прогнозована динаміка зростання рівня задоволеності користувачів за умови впровадження мультимодальної взаємодії

Розробка та інтеграція модуля сентимент-аналізу

Аналіз емоційного стану (сентимент-аналіз) виступає фундаментальним компонентом сучасних інтелектуальних систем обслуговування, забезпечуючи природність та гуманізацію діалогу. Здатність системи ідентифікувати емоційний фон користувача дозволяє динамічно адаптувати стратегії взаємодії, що безпосередньо впливає на покращення користувацького досвіду та підвищення рівня задоволеності. Попри наявність у системі Alibaba Xiaomi базових інструментів емоційного аналізу, існують значні резерви для зростання точності та глибини контекстуального розуміння, особливо в частині обробки складних емоційних патернів та мультимодальних даних. На поточному етапі функціонал Alibaba Xiaomi фокусується переважно на текстовій модальності, демонструючи обмежені можливості в інтерпретації інтонаційних нюансів мовлення, контекстуальних натяків та емодзі. Показовим є приклад, коли на фразу «Чудово! Нарешті знайшов вашу підтримку!» система може не розрізнити щиру радість від сарказму, що свідчить про слабкість алгоритмів у роботі з іронією, метафорами та амбівалентними емоціями. Крім того, відсутність механізму адаптації діалогових стратегій призводить до того, що навіть при виявленні роздратування клієнта система продовжує використовувати скриптовані, формалізовані відповіді.

Для подолання означених недоліків пропонується комплексна стратегія розвитку модуля сентимент-аналізу за трьома векторами. Перший вектор передбачає впровадження мультимодальної моделі аналізу настроїв, яка синтезує інформацію з тексту, аудіо та візуальних каналів для підвищення точності розпізнавання через злиття кросмодальних ознак. Застосування архітектури Transformer дозволить екстрагувати емоційні сигнали з інтонації, текстового контенту та смайлів, використовуючи спільний простір вбудовування (embedding space) та механізми уваги для фокусування на ключових маркерах. Наприклад, якщо клієнт голосовим повідомленням каже: «Ваш сервіс просто залишає мене без слів» і додає гнівний емодзі, система завдяки аналізу тону та символів зможе коректно ідентифікувати невдоволення та змінити стиль комунікації.

Другий вектор полягає у створенні механізму динамічного емоційного зворотного зв'язку, що дозволить коригувати діалогову стратегію в реальному часі на основі результатів аналізу. Це передбачає створення бібліотеки попередньо визначених сценаріїв реагування для різних емоційних станів: для позитивно налаштованих користувачів – заохочувальна риторика та пропозиція додаткових послуг; для негативно налаштованих – емпатична, заспокійлива лексика та пріоритетне вирішення проблеми. Наприклад, у відповідь на гнівний запит щодо повернення товару система має першочергово вибачитися та запевнити у швидкому вирішенні питання, замість стандартного пояснення бюрократичних процедур. Третій вектор спрямований на поглиблення розуміння складних емоційних контекстів (іронії, сарказму) шляхом врахування історичних даних діалогу через модуль відстеження контексту. Використання моделей глибокого навчання, таких як RoBERTa, донавчених на специфічних датасетах емоцій, дозволить системі здійснювати точніші судження на основі поточного вводу та попередньої історії взаємодії. Очікувана ефективність від впровадження вдосконаленого аналізу емоцій візуалізована на рисунку 5.

Реалізація запропонованих заходів створить підґрунтя для якісного покращення користувацького досвіду. Впровадження механізмів високоточного розпізнавання емоцій та динамічної адаптації діалогу дозволить системі демонструвати вищий рівень емпатії. Оптимізація стратегій роботи з негативним емоційним фоном сприятиме зниженню рівня скарг, тоді як ефективне реагування на позитивні емоції стимулюватиме зростання конверсійних показників. Особливого значення ці вдосконалення набувають у сферах з високими вимогами до емоційного

інтелекту – охороні здоров'я, освіті та фінансах, що дозволить суттєво розширити горизонти застосування системи Alibaba Xiaomi. Прогнозується, що оптимізація модуля сентимент-аналізу забезпечить приріст показника задоволеності користувачів на рівні 2–6%, зміцнюючи лідерські позиції платформи в сегменті інтелектуального обслуговування.

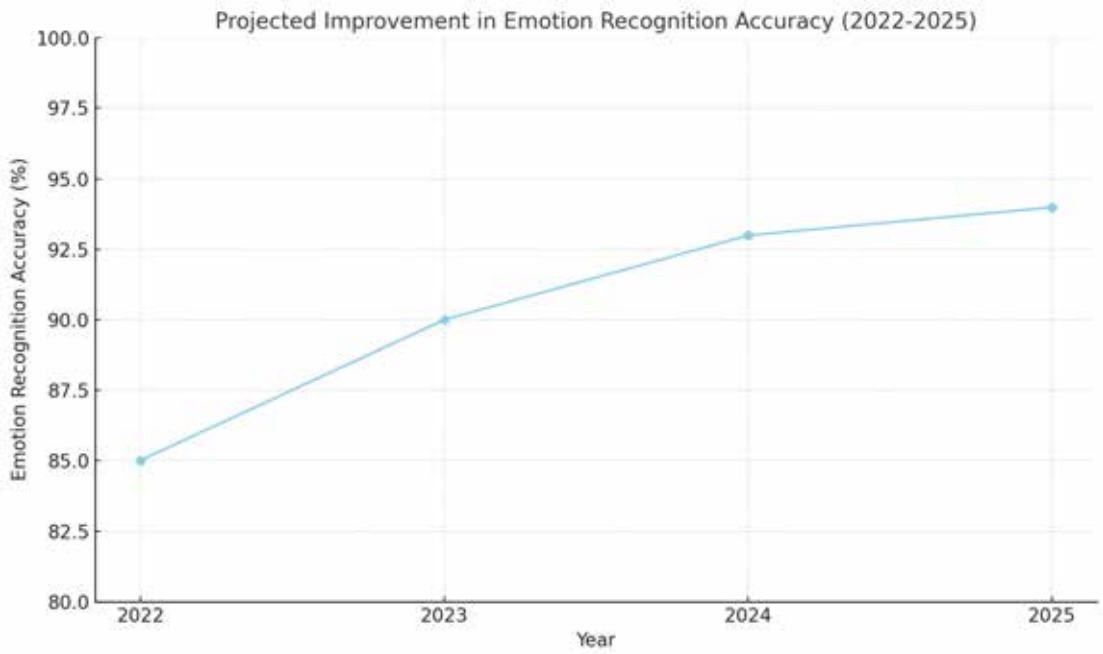

Рис. 5. Прогнозована динаміка підвищення точності розпізнавання емоцій (2022–2025 рр.)

Підсумовуючи, можна стверджувати, що стратегія розвитку мультимодального модуля аналізу емоцій, у поєднанні з механізмами динамічного зворотного зв'язку та поглибленим розумінням складних емоційних конструктів, дозволить Alibaba Xiaomi досягти принципово нового рівня природності та клієнтоорієнтованості сервісу. Така трансформація не лише радикально покращить сприйняття системи користувачами, а й генеруватиме додаткову вартість для бізнесу, формуючи одну з ключових конкурентних переваг сучасних інтелектуальних систем обслуговування.

Висновки

Здійснено комплексний аналіз пропозицій щодо оптимізації та тенденцій майбутнього розвитку інтелектуальних систем обслуговування клієнтів, а також запропоновано практичні рішення, що базуються на синергії технологічних інновацій та актуальних галузевих трендів. Для подолання функціональних обмежень існуючих систем сформульовано три ключові напрями вдосконалення: посилення можливостей управління контекстом для підвищення логічної зв'язності та результативності багатоходових діалогів; реалізація мультимодальної взаємодії через інтеграцію голосових, візуальних та текстових даних для роботи зі складними сценаріями; а також розробка модуля сентимент-аналізу, спрямованого на забезпечення більш природного, емпатичного та клієнтоорієнтованого сервісу.

Окрім тактичних кроків з оптимізації, окреслено стратегічні перспективи розвитку інтелектуальних систем. Ключовими векторами визначено глибоку інтеграцію з графами знань, що посилить здатність системи до логічних умовиводів та підвищить фаховий рівень відповідей, а також експансію у різноманітні галузеві сценарії (електронна комерція, фінанси, охорона здоров'я) для задоволення диференційованих потреб ринку. Окрему увагу приділено питанням вдосконалення механізмів захисту конфіденційності та забезпечення відповідності нормативним вимогам інформаційної безпеки. Поєднуючи теоретичний базис із аналізом практичних даних, цей розділ не лише визначає чіткий шлях технічної модернізації інтелектуальних систем обслуговування, а й формує стратегічний план їх подальшого масштабування на глобальному ринку.

Список використаної літератури

1. Yang Y. The Application of Natural Language Processing Technology in Legal Aid and Judicial Practice. *Academic Journal of Science and Technology*. Vol. 19, no. 2. 2026. Pp. 208-213. DOI: https://doi.org/10.54097/xvsq2r56

2. Farid A., Joel M., Gianluca D. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *ACM Comput. Surv*. Vol. 58, no. 6. 2025. P. 37. DOI: https://doi.org/10.1145/3777009

3. Song D, Gao S, He B, et al. On the effectiveness of pre-trained language models for legal natural language processing: An empirical study. *IEEE Access*. Vol. 10. 2022. Pp. 75835-75858. DOI: https://doi.org/10.1109/ACCESS.2022.3190408

4. Yang M. Design and implementation of intelligent customer service software architecture based on AI[J]. *Information recording materials*. Vol. 25, no. 5. 2024. Pp. 145-147. DOI: https://doi.org/10.16009/j.cnki.cn13-1295/TQ.2024.05.010

5. Barkovska O. Двофакторна автентифікація на основі методу KWS та голосової верифікації. *Сучасний стан наукових досліджень та технологій в промисловості*. Vol. 33, no. 3. 2025. Pp. 5-18. DOI: https://doi.org/10.30837/2522-9818.2025.3.005

6. Fedorchenko V., Yeroshenko O., Shmatko O., Kolomiitsev O., Omarov M. Password hashing methods and algorithms on the.Net platform, *Advanced Information Systems*. Vol. 8, no. 4. 2024. Pp. 82–92. DOI: https://doi.org/10.20998/2522-9052.2024.4.11

7. Alshebami A. S. Green innovation, self-efficacy, entrepreneurial orientation and economic performance: Interactions among Saudi small enterprises. *Sustainability*. Vol. 15, no. 3. 2023. Pp. 1961. DOI: https://doi.org/10.3390/su15031961

8. Ameen N., Sharma G. D., Tarba S., Rao A., Chopra R. Toward advancing theory on creativity in marketing and artificial intelligence. *Psychology and Marketing*. Vol. 39, no. 9. 2022. Pp. 180-1825. DOI: https://doi.org/10.1002/mar.21699Ashfaq

9. Su W, Hu Y, Xie A, et al. STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024. Pp. 10658-10671. DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.625

10. Lam J, Chen Y, Zulkernine F, et al. Legal text analytics for reasonable notice period prediction. *Journal of Computational and Cognitive Engineering*. 2025. DOI: https://doi.org/10.47852/bonviewJCCE52024104

11. Trautmann D, Petrova A, Schilder F. Legal prompt engineering for multilingual legal judgement prediction. *Computation and Language*. 2022. DOI: https://doi.org/10.48550/arXiv.2212.02199

12. Quevedo E, Cerny T, Rodriguez A, et al. Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access*. Vol. 12. 2023. Pp. 145286-145317. DOI: https://doi.org/10.1109/ACCESS.2023.3333946

13. Barkovska O., Ruban I., Tymoshenko D., Holovchenko O., Yankovskyi O. Research on mobile machine learning platforms for human gesture recognition in human-machine interaction systems. *Technology Audit and Production Reserves.* Vol. 82, no. 2. 2025. Pp. 6–14. DOI: https://doi.org/10.15587/2706-5448.2025.325423

14. Yeroshenko O., Andrushchenko A., Sorokin A., Sitnikov V. Application of ensemble learning algorithms for fraud detection in banking transactions, *Вісник Херсонського національного технічного університету*. Vol. 94, no. 3. 2025. Pp. 186-191. DOI: https://doi.org/10.35546/kntu2078-4481.2025.3.2.23

15. Федорченко В. М., Єрошенко О. А. Застосування алгоритмів штучного інтелекту для моделювання загроз інформаційних систем, *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки»*. Vol. 75, no. 6. 2025. Pp. 384-391. DOI: https://doi.org/10.32782/2663-5941/2025.6.2/52

References

1. Yang, Y. (2026) The Application of Natural Language Processing Technology in Legal Aid and Judicial Practice. *Academic Journal of Science and Technology*, vol. 19, no. 2, pp. 208-213. DOI: https://doi.org/10.54097/xvsq2r56

2. Farid, A., Joel, M., Gianluca, D. (2025) Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges. *ACM Comput. Surv*. vol. 58, no. 6, p. 37. DOI: https://doi.org/10.1145/3777009

3. Song, D, Gao, S, He, B, et al. (2022) On the effectiveness of pre-trained language models for legal natural language processing: An empirical study. *IEEE Access*, vol. 10, pp. 75835-75858. DOI: https://doi.org/10.1109/ACCESS.2022.3190408

4. Yang, M. (2024) Design and implementation of intelligent customer service software architecture based on AI[J]. *Information recording materials*, vol. 25, no. 5, pp. 145-147. DOI: https://doi.org/10.16009/j.cnki.cn13-1295/TQ.2024.05.010

5. Barkovska, O. (2025) Two-factor authentication based on keyword spotting and speaker verification. *Innovative Technologies and Scientific Solutions for Industries*, vol. 33. no. 3, pp. 5-18. DOI: https://doi.org/10.30837/2522-9818.2025.3.005

6. Fedorchenko, V., Yeroshenko, O., Shmatko, O., Kolomiitsev, O., Omarov, M. (2024) Password hashing methods and algorithms on the.Net platform, *Advanced Information Systems*, vol. 8, no. 4, pp. 82–92. DOI: https://doi.org/10.20998/2522-9052.2024.4.11

7. Alshebami, A. S. (2023) Green innovation, self-efficacy, entrepreneurial orientation and economic performance: Interactions among Saudi small enterprises. *Sustainability*, vol. 15, no. 3, pp. 1961. DOI: https://doi.org/10.3390/su15031961

8. Ameen, N., Sharma, G. D., Tarba, S., Rao, A., Chopra, R. (2022) Toward advancing theory on creativity in marketing and artificial intelligence. *Psychology and Marketing*, vol. 39, no. 9, pp. 180-1825. DOI: https://doi.org/10.1002/mar.21699Ashfaq

9. Su, W, Hu, Y, Xie, A, et al. (2024) STARD: A Chinese statute retrieval dataset derived from real-life queries by non-professionals. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10658-10671. DOI: https://doi.org/10.18653/v1/2024.findings-emnlp.625

10. Lam, J, Chen, Y, Zulkernine, F, et al. (2025) Legal text analytics for reasonable notice period prediction. *Journal of Computational and Cognitive Engineering*. DOI: https://doi.org/10.47852/bonviewJCCE52024104

11. Trautmann, D, Petrova, A, Schilder, F. (2022) Legal prompt engineering for multilingual legal judgement prediction. *Computation and Language*. DOI: https://doi.org/10.48550/arXiv.2212.02199

12. Quevedo, E, Cerny, T, Rodriguez, A, et al. (2023) Legal natural language processing from 2015 to 2022: A comprehensive systematic mapping study of advances and applications. *IEEE Access*, vol. 12, pp. 145286-145317. DOI: https://doi.org/10.1109/ACCESS.2023.3333946

13. Barkovska, O., Ruban, I., Tymoshenko, D., Holovchenko, O., Yankovskyi, O. (2025) Research on mobile machine learning platforms for human gesture recognition in human-machine interaction systems. *Technology Audit and Production Reserves*, vol. 82, no. 2, pp. 6–14. DOI: https://doi.org/10.15587/2706-5448.2025.325423

14. Yeroshenko, O., Andrushchenko, A., Sorokin, A., Sitnikov, V. (2025) Application of ensemble learning algorithms for fraud detection in banking transactions, *Вісник Херсонського національного технічного університету*, vol. 94, no. 3, pp. 186-191. DOI: https://doi.org/10.35546/kntu2078-4481.2025.3.2.23

15. Fedorchenko, V., Yeroshenko, O. (2025) Zastosuvannya alhorytmiv shtuchnoho intelektu dlya modelyuvannya zahroz informatsiynykh system [Application of artificial intelligence algorithms for modeling threats to information systems], *Scientific notes of the V.I. Vernadsky Tavrichesky National University. Series: Technical Sciences*, vol. 75, no. 6, pp. 384-391. DOI: https://doi.org/10.32782/2663-5941/2025.6.2/52

Дата першого надходження статті до видання: 26.02.2026
Дата прийняття статті до друку після рецензування: 30.03.2026
Дата публікації (оприлюднення) статті: 07.05.2026