№ 3(57)
2026
НАУКА і ТЕХНІКА СЬОГОДНІ
ISSN 2786-6025 Online
УДК 004:089

https://doi.org/10.52058/2786-6025-2026-3(57)-3190-3200

Яровий Андрій Анатолійович доктор технічних наук, професор, завідувач кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, https://orcid.org/0000-0002-6668-2425

Кудрявцев Дмитро Станіславович доктор філософії, асистент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, https://orcid.org/0000-0001-7116-7869

Петришин Сергій Іванович кандидат технічних наук, старший викладач кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, https://orcid.org/0009-0001-3465-1499

Озеранський Володимир Сергійович кандидат технічних наук, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, https://orcid.org/0009-0007-1694-2317

Ваховська Любов Михайлівна асистент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця, https://orcid.org/0000-0002-4865-6514

БАГАТОШАРОВА НЕЙРОННА АРХІТЕКТУРА ІНТЕЛЕКТУАЛЬНОГО ЧАТ-БОТА НА ОСНОВІ КОМБІНАЦІЇ МОДЕЛІ ТРАНСФОРМЕРА ТА ТЕРМІНОЛОГІЧНИХ БАЗ ЗНАНЬ

Анотація. У статті представлено концепцію та формалізацію багатошарової нейронної архітектури інтелектуального чат-бота, орієнтованого на одночасну обробку декількох предметних областей із можливістю динамічного уточнення та розширення контексту запиту користувача. Актуальність дослідження зумовлена обмеженнями універсальних великих мовних моделей, які, забезпечуючи високий рівень генерації тексту, демонструють недостатню глибину спеціалізації та знижену точність у вузькопрофесійних або міждисциплінарних задачах.

Запропонований підхід базується на інтеграції трансформерних моделей глибинного навчання, рекурентних нейронних мереж із використанням методу LSTM та структурованих термінологічних баз знань у межах гібридної інформаційної технології.

ISSN 2786-6025 Online

Архітектура передбачає багаторівневу обробку текстового сигналу, що включає первинну NLP-нормалізацію, формування контекстних векторних представлень, ймовірнісну доменну класифікацію, кластеризацію термів та їх семантичне зіставлення зі структурованими базами знань. Особливістю моделі є реалізація механізму мультидоменної маршрутизації, який дозволяє одночасно активувати декілька предметних областей та здійснювати адаптивний розподіл ваг між ними. Це забезпечує коректну інтерпретацію міждисциплінарних запитів і зменшує ризик втрати семантичної точності.

Центральним елементом запропонованого підходу є метод багатоцільового пошуку, що передбачає багатоступеневу фільтрацію релевантних термів та динамічне розширення семантичного ядра запиту за рахунок споріднених понять. Такий механізм дозволяє зменшити семантичний шум, локалізувати пошук у межах релевантних предметних областей та оптимізувати обчислювальні витрати. Генеративний контур реалізовано на основі поєднання моделі трансформера, відповідальної за глобальну семантичну узгодженість відповіді, та LSTM-модуля, що забезпечує збереження локального діалогового контексту і когерентність взаємодії.

Запропонована архітектура дозволяє підвищити точність визначення контексту, покращити деталізацію відповідей у вузькоспеціалізованих сферах та забезпечити структуровану мультидоменну обробку порівняно з класичними підходами.

Результати дослідження можуть бути використані при розробці інтелектуальних інформаційних систем нового покоління, орієнтованих на персоналізовану взаємодію з користувачем.

Ключові слова: алгоритмізація, програмування, LSTM, ООП, база знань, нейронна мережа, LLM, інтелектуальна інформаційна технологія, гібридні обчислювальні системи, штучний інтелект.

Yarovyi Andrii Anatoliyovych Doctor of Engineering Sciences, Professor, Head of the Department for Computer Science, Vinnytsia National Technical University, Vinnytsia, https://orcid.org/0000-0002-6668-2425

Kudriavtsev Dmytro Stanislavovych PhD, Assistant of the Department of Computer Science, Vinnytsia National Technical University, Vinnytsia, https://orcid.org/0000-0001-7116-7869

Petrishyn Sergiy Ivanovych Candidate of Engineering Sciences, Senior Lecturer at the Department of Computer Science, Vinnytsia National Technical University, 95 Khmelnytske shose, Vinnytsia, https://orcid.org/0009-0001-3465-1499

ISSN 2786-6025 Online

Ozeranskyi Volodymyr Serhiiovych Candidate of Engineering Sciences, Associate Professor at the Department for Computer Science, Vinnytsia National Technical University, Vinnytsia, https://orcid.org/0009-0007-1694-2317

Vahovska Lyubov Mykhailivna Assistant at the Department for Computer Science, Vinnytsia National Technical University, Vinnytsia, https://orcid.org/0000-0002-4865-6514

MULTILAYER NEURAL ARCHITECTURE OF AN INTELLIGENT CHATBOT BASED ON THE COMBINATION OF A TRANSFORMER MODEL AND TERMINOLOGICAL KNOWLEDGE BASES

Abstract. The article presents the concept and formalization of a multilayer neural architecture of an intelligent chatbot designed for simultaneous processing of multiple subject domains with the capability of dynamic clarification and expansion of the user query context. The relevance of the study is determined by the limitations of universal large language models which, while providing a high level of text generation, demonstrate insufficient depth of specialization and reduced accuracy in narrow professional or interdisciplinary tasks. The proposed approach is based on the integration of transformer-based deep learning models, recurrent neural networks of the LSTM type, and structured terminological knowledge bases within a hybrid information technology framework.

The architecture involves multi-level processing of the textual signal, including primary NLP normalization, formation of contextual vector representations, probabilistic domain classification, term clustering, and semantic matching with structured knowledge bases. A distinctive feature of the model is the implementation of a multi-domain routing mechanism that enables simultaneous activation of multiple subject domains and adaptive weight distribution among them. This ensures correct interpretation of interdisciplinary queries and reduces the risk of semantic precision loss.

The central element of the proposed approach is a multi-objective search method that includes multi-stage filtering of relevant terms and dynamic expansion of the semantic core of the query through related concepts. Such a mechanism allows reducing semantic noise, localizing the search within relevant subject domains, and optimizing computational costs. The generative pathway is implemented through the combination of a transformer model responsible for global semantic coherence of the response and an LSTM module that ensures preservation of local dialogue context and interaction coherence.

The proposed architecture makes it possible to increase context determination accuracy, improve response detailing in narrowly specialized domains, and provide

ISSN 2786-6025 Online
structured multi-domain processing compared to classical approaches. The results of the study can be applied in the development of next-generation intelligent information systems focused on personalized and professionally oriented user interaction
Keywords: algorithm, programming, LSTM, OOP, knowledge base, neural network, LLM, intelligent information technology, hybrid computing systems, arteficial intelligence.

Постановка проблеми. Стрімкий розвиток технологій штучного інтелекту, зокрема архітектури трансформера та великих мовних моделей (LLM), суттєво змінив підхід до побудови систем обробки природної мови. Сучасні чат-боти демонструють високий рівень генерації тексту, здатність підтримувати діалог, аналізувати складні синтаксичні структури та формувати відповіді з урахуванням довгих контекстних залежностей. Проте, попри очевидні досягнення, універсальні мовні моделі мають низку системних обмежень, які стають особливо помітними у предметно-орієнтованих або міждисциплінарних задачах.

По-перше, більшість великих мовних моделей формують відповіді на основі ймовірнісного прогнозування наступних токенів. Такий підхід забезпечує узагальнену семантичну релевантність, однак не гарантує структурованої відповідності конкретній предметній області. В умовах вузькопрофесійного запиту це може призводити до поверхневих або частково коректних відповідей, які не враховують специфіку термінології, ієрархію понять та логіку предметної області [1].

По-друге, універсальність LLM передбачає одночасне охоплення великої кількості доменів, що неминуче знижує рівень спеціалізації кожного з них. У результаті модель оперує усередненими семантичними представленнями, що не завжди достатньо точні для задач, які потребують глибокої деталізації або інтеграції декількох предметних сфер.

Особливо це проявляється у міждисциплінарних запитах, де модель змушена одночасно враховувати декілька контекстних площин без наявності явного механізму їх структурованої координації [1].

По-третє, існуючі архітектури переважно використовують єдиний контекстний простір без формалізованої доменної маршрутизації. Відсутність механізму адаптивного розподілу ваг між предметними областями унеможливлює коректну інтерпретацію запиту. У результаті система змушена здійснювати жорстку класифікацію або покладатися на статистичний компроміс, що знижує точність визначення релевантного контексту.

Крім того, проблема ускладнюється відсутністю структурованого механізму контролю семантичного ядра запиту. Більшість моделей не виконують явного розширення або фільтрації термінологічного простору, що призводить

ISSN 2786-6025 Online

до накопичення семантичного шуму. У випадку спеціалізованих задач це може викликати часткову втрату змістової точності або генерацію відповіді з надлишковою узагальненістю [1].

Таким чином, актуальною є задача побудови такої архітектури інтелектуального чат-бота, яка б поєднувала глибинні нейронні моделі з формалізованими термінологічними базами знань та забезпечувала багаторівневе визначення контексту. Особливого значення набуває необхідність реалізації механізму мультидоменного аналізу, що дозволяє одночасно активувати декілька предметних областей, здійснювати контрольоване розширення семантичного ядра запиту та мінімізувати семантичний шум.

Отже, проблема полягає у відсутності інтегрованої багатошарової архітектури, яка б забезпечувала синергію між моделями глибинного навчання, рекурентними механізмами збереження локального контексту та структурованими базами знань у межах єдиної інформаційної технології. Розв'язання цієї проблеми дозволить підвищити точність визначення контексту, покращити деталізацію відповідей у вузькоспеціалізованих сферах та забезпечити адаптивність системи до складних міждисциплінарних запитів.

Аналіз останніх досліджень і публікацій. Сучасні дослідження у сфері інтелектуальних чат-ботів та систем обробки природної мови концентруються переважно навколо розвитку трансформерних архітектур та масштабування великих мовних моделей (LLM). Моделі типу GPT, BERT та їх похідні стали базовим інструментом для генерації тексту, аналізу контексту та підтримки діалогу. Основною перевагою таких моделей є здатність формувати контекстні представлення довгих послідовностей та моделювати складні семантичні залежності за допомогою механізму self-attention [2].

Разом із тим у міжнародних публікаціях дедалі частіше наголошується, що масштабування параметрів моделі саме по собі не гарантує підвищення точності у вузькоспеціалізованих або міждисциплінарних задачах.

Універсальні мовні моделі демонструють статистичну узгодженість відповідей, однак не завжди забезпечують структуровану відповідність конкретній предметній області. Це особливо актуально для запитів, що потребують точного термінологічного аналізу або інтеграції декількох доменів одночасно.

У відповідь на ці обмеження активно розвивається напрям Retrieval-Augmented Generation (RAG), який передбачає поєднання генеративної моделі з механізмом пошуку релевантного контексту у зовнішніх джерелах знань [3,4]. Подальшим розвитком цієї концепції є використання графів знань (knowledge graphs), що дозволяє враховувати ієрархічні та асоціативні зв'язки між поняттями. Такі підходи демонструють покращення точності відповідей у задачах, де важливою є формалізована структура знань.

ISSN 2786-6025 Online

Окремий напрям досліджень пов'язаний із мультидоменними діалоговими системами. У сучасних роботах пропонуються механізми доменної маршрутизації, використанню вагів та багаторівневої уваги, які дозволяють підсилювати релевантні стани пам`яті та зменшувати вплив нерелевантної інформації. Зазначається, що жорстка класифікація запиту до одного домену часто є недостатньою, особливо у випадку міждисциплінарних задач, що потребують одночасної активації декількох предметних областей [4, 5].

Крім того, у низці робіт розглядаються гібридні архітектури, які поєднують трансформерні моделі з рекурентними мережами або зовнішніми модулями керування контекстом. Такий підхід дозволяє розмежувати глобальне семантичне моделювання та локальну діалогову динаміку, що позитивно впливає на когерентність відповіді в межах сесії взаємодії [6,7].

Аналіз міжнародних публікацій свідчить про поступовий перехід від суто генеративних моделей до гібридних систем, що інтегрують нейронні мережі з формалізованими структурами знань та механізмами адаптивної маршрутизації. Водночас питання контрольованого розширення семантичного ядра запиту, багатоступеневої фільтрації термінів і одночасної активації залишаються недостатньо формалізованими в межах єдиної архітектурної концепції [1, 7]. Саме це визначає доцільність розробки багатошарової гібридної моделі, що поєднує аналіз, багатоцільовий пошук у термінологічних базах знань та механізм маршрутизації в різних контекстах.

Мета статті – дослідження та формалізація багатошарової нейронної архітектури інтелектуального чат-бота, що поєднує модель трансформера, рекурентний модуль із використанням методу LSTM та термінологічні бази знань для забезпечення мультидоменного визначення контексту та багатоцільового семантичного пошуку.

Виклад основного матеріалу. Запропонована багатошарова нейронна архітектура інтелектуального чат-бота розглядається як гібридна інформаційна система, що поєднує статистичні методи глибинного навчання із структурованими термінологічними базами знань та механізмами мультидоменного керування контекстом. Концептуально архітектура реалізує принцип поступового уточнення семантичного представлення запиту користувача, де кожен наступний шар виконує функцію деталізації та зменшення семантичної невизначеності.

ISSN 2786-6025 Online

Рис. 1. - Схема початкової обробки повідомлення користувача

Початковим етапом є первинна обробка тексту, яка виконує роль ентропійного зниження мовного сигналу. У цьому контурі реалізуються операції токенізації, лематизації, морфологічного аналізу та виділення іменованих сутностей.

Послідовність цих процедур зумовлена необхідністю нормалізації словоформ і мінімізації синтаксичної варіативності перед формуванням векторного представлення [1].

Після нормалізації текст передається до модуля трансформера, який формує контекстні зв`язки із використанням механізму self-attention. Важливо підкреслити, що трансформер у запропонованій архітектурі не є остаточним інструментом прийняття рішення, а виступає як шар формування багатовимірного семантичного простору, який слугує основою для подальшої структурованої обробки.

№ 3(57) 2026
НАУКА і ТЕХНІКА СЬОГОДНІ
ISSN 2786-6025 Online

Рис. 2. - Схема визначення контексу повідомлення користувача

Наступним етапом є багаторівневе визначення контексту запиту. Цей модуль реалізує інтегроване застосування ймовірнісної класифікації та кластеризації термів. На першому рівні формується розподіл імовірностей належності запиту до визначених предметних областей. На відміну від класичної жорсткої класифікації, модель зберігає весь спектр доменних ваг, що дозволяє уникнути передчасної редукції контексту. На другому рівні виконується кластеризація активованих термів у векторному просторі для виявлення латентних семантичних груп. Це дозволяє виявити приховані тематичні зв’язки, які можуть бути неочевидними на рівні поверхневої лексики.

На третьому рівні здійснюється зіставлення отриманих термів із структурованими термінологічними базами знань. Саме на цьому етапі відбувається інтеграція статистичної семантики з формалізованими зв’язками між поняттями, що забезпечує підвищення точності визначення релевантного контексту. Особливу роль у архітектурі відіграє механізм мультидоменної маршрутизації. У запропонованій моделі запит може одночасно активувати декілька предметних областей, якщо інтегральна функція оцінки перевищує встановлений поріг релевантності. Ваги доменів визначаються з урахуванням ймовірнісної класифікації, щільності відповідних термів у базах знань та коефіцієнта семантичної близькості між доменами. Такий підхід дозволяє коректно інтерпретувати міждисциплінарні запити, уникати семантичної редукції та забезпечувати адаптивний розподіл обчислювальних ресурсів.

ISSN 2786-6025 Online

Ключовим елементом системи є метод багатоцільового пошуку, що реалізує багатоступеневу фільтрацію та контрольоване розширення семантичного ядра запиту [1]. На першому етапі здійснюється прямий пошук співпадінь у відповідних термінологічних базах знань. На другому етапі відбувається розширення ядра через включення споріднених термів на основі ієрархічних, асоціативних та статистичних зв'язків. Це дозволяє враховувати латентні семантичні залежності та підвищувати повноту аналізу. На третьому етапі виконується багатоступенева фільтрація, яка передбачає відсіювання термів за порогами семантичної релевантності та структурної значущості.

Таким чином формується уточнене семантичне ядро, що зменшує шум і локалізує пошук у межах релевантних доменів. Генеративний контур архітектури побудований на поєднанні трансформерного модуля та рекурентної нейронної мережі з використанням методу LSTM. Трансформер відповідає за глобальну семантичну узгодженість відповіді, враховуючи розширене семантичне ядро та ваги активованих доменів. Його механізм уваги дозволяє інтегрувати інформацію з різних предметних областей у єдину відповідь. LSTM-модуль виконує функцію збереження локального діалогового контексту, забезпечуючи когерентність, логічну послідовність та персоналізацію взаємодії в межах сесії.

Поєднання цих компонентів дозволяє досягти балансу між глобальним моделюванням знань та локальною динамікою діалогу.

У результаті запропонована архітектура формує замкнений цикл обробки: від первинного семантичного представлення до структурованого доменного уточнення та адаптивної генерації відповіді. Такий підхід дозволяє підвищити точність визначення контексту, зменшити семантичний шум, забезпечити мультидоменну інтерпретацію запитів та оптимізувати обчислювальні витрати порівняно з універсальними генеративними моделями.

Висновки. У статті досліджено та теоретично обґрунтовано багатошарову нейронну архітектуру інтелектуального чат-бота, що поєднує трансформерну модель глибинного навчання, рекурентний модуль із використанням методу LSTM та структуровані термінологічні бази знань у межах єдиної гібридної інформаційної технології. Запропонований підхід орієнтований на розв'язання задачі мультидоменної обробки запитів користувача з динамічним уточненням контексту та контрольованим розширенням семантичного ядра.

У ході дослідження встановлено, що використання трансформерної моделі як інструменту формування контекстного векторного простору є ефективним лише за умови подальшої структурованої інтерпретації отриманих embeddings. Саме інтеграція статистичного семантичного представлення з формалізованими термінологічними базами знань дозволяє підвищити точність

ISSN 2786-6025 Online

визначення релевантного доменного контексту та мінімізувати вплив семантичного шуму.

Запропонований механізм багаторівневого визначення контексту, який поєднує ймовірнісну класифікацію, кластеризацію термів та їх структуроване зіставлення з базами знань, забезпечує більш глибоку інтерпретацію запитів порівняно з традиційною жорсткою доменною класифікацією. Реалізація мультидоменної маршрутизації дозволяє одночасно активувати декілька предметних областей та адаптивно розподіляти ваги між ними, що є принципово важливим для міждисциплінарних задач.

Метод багатоцільового пошуку з багатоступеневою фільтрацією та контрольованим розширенням семантичного ядра продемонстрував здатність локалізувати обчислювальний процес у межах релевантних доменів, зменшити надлишковість термінів і підвищити деталізацію відповідей у вузькоспеціалізованих сферах. Поєднання трансформерного генеративного модуля з LSTM-компонентом дозволило забезпечити баланс між глобальною семантичною узгодженістю відповіді та локальною діалоговою когерентністю в межах сесії взаємодії.

Таким чином, запропонована архітектура реалізує концепцію гібридної мультидоменної діалогової системи, яка перевершує класичні універсальні LLM-підходи за критеріями контекстної точності, адаптивності та керованості семантичного простору. Практичне значення результатів полягає у можливості використання розробленої моделі для створення професійно орієнтованих чатботів у галузях освіти, інженерії, програмування, маркетингу та інших сферах, де необхідна точна термінологічна інтерпретація запитів.

Перспективи подальших досліджень пов’язані з формалізацією механізму автоматичного формування нових предметних областей, оптимізацією міжпредметного механізму уваги та впровадженням елементів навчання для адаптивного налаштування ваг маршрутизації в умовах динамічного оновлення баз знань.

Література:

1. A. Yarovyi and D. Kudriavtsev, “FORMATION OF HIGHLY SPECIALIZED CHATBOTS FOR ADVANCED SEARCH”, IAPGOS, vol. 14, no. 1, pp. 67–70, Mar. 2024. Режим доступу: http://dx.doi.org/10.35784/iapgos.5628

2. Singh S. U. *A comprehensive survey on chatbots and large language models: testing, evaluation and performance*. Journal of Applied Artificial Intelligence Research, 2025. Режим доступу: https://www.sciencedirect.com

3. Yang W. A comprehensive survey on integrating large language models and knowledge bases. *Knowledge-Based Systems*, 2025. Режим доступу: https://www.sciencedirect.com

4. Mienye I. D. Large language models: architecture, training methodologies and emerging trends. *SN Applied Sciences*, 2025. Режим доступу: https://link.springer.com

ISSN 2786-6025 Online

5. Hyder S. Designing intelligent chatbots with ChatGPT: a framework for hybrid model design and implementation. *Frontiers in Artificial Intelligence*, 2025. Режим доступу: https://www.frontiersin.org

6. Al-Shamaileh O. Navigating ethical considerations and implications of AI chatbots in higher education. *Computers & Education: Artificial Intelligence*, 2026. Режим доступу: https://www.sciencedirect.com

7. Kristiani E. Deploying LLM transformer on edge computing devices: strategies, challenges, future directions. *MDPI Intelligent Systems*, 2026. Режим доступу: https://www.mdpi.com

References:

1. A. Yarovyi and D. Kudriavtsev, "*FORMATION OF HIGHLY SPECIALIZED CHATBOTS FOR ADVANCED SEARCH*", IAPGOS, vol. 14, no. 1, pp. 67–70, Mar. 2024. Available at: http://dx.doi.org/10.35784/iapgos.5628 (accessed 18 February 2026).

2. Singh S. U. A comprehensive survey on chatbots and large language models: testing, evaluation and performance. *Journal of Applied Artificial Intelligence Research*, 2025. Available at: https://www.sciencedirect.com (accessed 18 February 2026).

3. Yang W. A comprehensive survey on integrating large language models and knowledge bases. *Knowledge-Based Systems*, 2025. Available at: https://www.sciencedirect.com (accessed 18 February 2026).

4. Mienye I. D. Large language models: architecture, training methodologies and emerging trends. *SN Applied Sciences*, 2025. Available at: https://link.springer.com (accessed 18 February 2026).

5. Hyder S. Designing intelligent chatbots with ChatGPT: a framework for hybrid model design and implementation. *Frontiers in Artificial Intelligence*, 2025. Available at: https://www.frontiersin.org (accessed 18 February 2026).

6. Al-Shamaileh O. Navigating ethical considerations and implications of AI chatbots in higher education. *Computers & Education: Artificial Intelligence*, 2026. Available at: https://www.sciencedirect.com (accessed 18 February 2026).

7. Kristiani E. Deploying LLM transformer on edge computing devices: strategies, challenges, future directions. *MDPI Intelligent Systems*, 2026. Available at: https://www.mdpi.com (accessed 18 February 2026).

Дата першого надходження статті до видання: 10.03.2026
Дата прийняття статті до друку після рецензування: 26.03.2026