УДК 004.896 004.415

С. О. Цимбал
Н. П. Бабюк

АРХІТЕКТУРА У СФЕРІ ШТУЧНОГО ІНТЕЛЕКТУ ТА МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Стаття присвячена дослідженню еволюції архітектурних підходів у сфері штучного інтелекту та машинного навчання. Розглядається перехід від традиційних монолітних моделей до складних агентних фреймворків та екосистем. Проаналізовано новітні рішення щодо відокремлення пам'яті від логічного висновку, розв'язання проблем масштабування та впровадження динамічної маршрутизації між різними мовними моделями. Значна увага приділяється інтеграції хмарних API для створення масштабованих проєктів та архітектурній адаптації до нових нормативних вимог.

Ключові слова: штучний інтелект, машинне навчання, архітектура ПЗ, LLM, маршрутизація моделей, агентні системи.

Abstract

The article investigates the evolution of architectural approaches in the field of artificial intelligence and machine learning. It examines the transition from traditional monolithic models to complex agentic frameworks and ecosystems. The latest solutions regarding the separation of memory from reasoning, solving scaling problems, and implementing dynamic routing between different language models are analyzed. Significant attention is paid to the integration of cloud APIs for creating scalable projects and architectural adaptation to new regulatory requirements.

Keywords: artificial intelligence, machine learning, software architecture, LLM, model routing, agentic systems.

Вступ

Стрімкий розвиток генеративного штучного інтелекту та великих мовних моделей кардинально змінив підходи до проєктування програмного забезпечення. Сучасні системи машинного навчання поступово відходять від парадигми ізольованих монолітних нейронних мереж і трансформуються у складні розподілені компаундні екосистеми. Ця еволюція зумовлена необхідністю оптимізації обчислювальних ресурсів, забезпечення жорстких вимог щодо безпеки та надійності, а також потребою в інтеграції різноманітних когнітивних компонентів у єдиний інтегрований конвеєр. Розробка архітектури штучного інтелекту сьогодні вимагає не лише глибокого розуміння математичних моделей, але й застосування передових практик програмної інженерії для керування даними, апаратною інфраструктурою та безперервним життєвим циклом розгортання моделей. Нові виклики вимагають створення адаптивних і модульних рішень, здатних швидко еволюціонувати відповідно до динамічних вимог індустрії та регуляторних норм.

Результати дослідження

Одним із головних трендів сучасного проєктування є використання агентних архітектур, де штучний інтелект виступає не лише як пасивний інструмент обробки даних, але й як активний учасник процесу оптимізації та пошуку рішень. Інтеграція агентних фреймворків дозволяє автоматизувати дослідження складних просторів проєктування, поєднуючи еволюцію коду на основі великих мовних моделей із симуляцією архітектурних змін у реальному часі [1]. Такий підхід суттєво змінює роль інженера, перетворюючи його на постановника завдань та контролера якості, тоді як рутинні операції з пошуку оптимальних мікро архітектурних рішень делегуються автономним агентам [2].

Паралельно з еволюцією організаційних підходів відбуваються фундаментальні зміни в апаратній та алгоритмічній архітектурі обслуговування самих моделей машинного навчання.

Важливим проривом є впровадження систем ефективного управління пам'яттю під час логічного висновку, які використовують механізми розподілу на основі сторінок для оптимізації пропускної здатності та паралельної генерації тексту [3]. Такі інновації безпосередньо впливають на економіку розробки та закони масштабування, дозволяючи розгортати моделі з екстремально довгим контекстом без експоненційного зростання витрат на дорогу інфраструктуру графічних процесорів [4].

Додатковим напрямком радикальної оптимізації обчислювальних ресурсів є масовий перехід індустрії до архітектури суміші експертів. Цей підхід передбачає створення розріджених нейронних мереж, де під час обробки кожного запиту активується лише невелика частина параметрів моделі, яка найкраще підходить для конкретного семантичного завдання [5]. Архітектура суміші експертів дозволяє значно збільшити загальну ємність бази знань системи без пропорційного збільшення часу на обробку одного токена, що є критично важливим для створення над швидких корпоративних рішень.

Невід'ємною складовою сучасних архітектур штучного інтелекту стала глибока інтеграція систем генерації, доповненої пошуком. Замість того, щоб зберігати всі фактологічні знання виключно у вагах нейронної мережі, розробники проєктують системи, які динамічно витягують релевантну інформацію із зовнішніх джерел перед формуванням фінальної відповіді [6]. Це архітектурне рішення не лише вирішує фундаментальну проблему галюцинацій мовних моделей, але й дозволяє оновлювати контекст у режимі реального часу. Процес об'єднання мовних моделей із системами швидкого доступу до знань вимагає побудови надійних конвеєрів обробки даних та механізмів високоточної семантичної індексації на базі спеціалізованих векторних сховищ [7].

Особливої актуальності набуває побудова хмарних архітектур з використанням патернів динамічної маршрутизації між різними мовними моделями. Сучасні платформи інтегрують ієрархічні маршрутизатори, які автоматично підбирають оптимальну за вартістю та якістю модель для кожного конкретного завдання [8]. Водночас зростаюча автономність таких систем вимагає впровадження окремих архітектурних шарів безпеки, так званих захисних бар'єрів, які безперервно фільтрують вхідні запити та згенеровані відповіді на відповідність етичним нормам [9]. Використання мікросервісних підходів для надійної оркестрації викликів до моделей та паралельних систем захисту дозволяє ефективно реалізовувати проєкти без необхідності підтримки нестабільних локальних кластерів обчислень [10].

Висновки

Сучасна архітектура у сфері штучного інтелекту та машинного навчання характеризується остаточним переходом до модульних агентних систем та розріджених мереж на основі концепції суміші експертів. Ключову роль відіграє оркестрація взаємодії між спеціалізованими компонентами, векторними базами даних та механізмами генерації, доповненої пошуком. Інновації на рівні ефективного управління пам'яттю під час логічного висновку та впровадження інтелектуальної маршрутизації запитів суттєво підвищують продуктивність і знижують загальну вартість експлуатації цифрових продуктів. Інтеграція хмарних рішень спільно з надійними архітектурними шарами безпеки дозволяє інженерам створювати складні, масштабовані та відмовостійкі проєкти, повністю нівелюючи попередні інфраструктурні обмеження. Подальший розвиток галузі вимагатиме суворої стандартизації механізмів інтеграції та адаптації до нових глобальних регуляторних стандартів надійності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1.arXiv. A Survey on Large Language Model based Autonomous Agents. URL: https://arxiv.org/abs/2308.11432 (дата звернення: 02.06.2026).

2.arXiv. LLM as OS, Agents as Apps: Envisioning AIOS, Agents and the AIOS-Agent Ecosystem. URL: https://arxiv.org/abs/2403.16971 (дата звернення: 02.06.2026).

3.ACM Digital Library. Efficient Memory Management for Large Language Model Serving with PagedAttention. URL: https://dl.acm.org/doi/10.1145/3600006.3613165 (дата звернення: 02.06.2026).

4.arXiv. LLM Architecture, Scaling Laws, and Economics: A Quick Summary. URL: https://arxiv.org/abs/2511.11572 (дата звернення: 02.06.2026).

5.arXiv. Mixtral of Experts (Sparse Mixture-of-Experts language model). URL: https://arxiv.org/abs/2401.04088 (дата звернення: 02.06.2026).

6.ACM Computing Surveys. A Survey on Retrieval-Augmented Text Generation for Large Language Models. URL: https://dl.acm.org/doi/10.1145/3805774 (дата звернення: 02.06.2026).

7.ACM Transactions on Information Systems. Graph Retrieval-Augmented Generation: A Survey. URL: https://dl.acm.org/doi/10.1145/3777378 (дата звернення: 02.06.2026).

8.LMSYS Blog. RouteLLM: An Open-Source Framework for Cost-Effective LLM Routing. URL: https://www.lmsys.org/blog/2024-07-01-routellm/ (дата звернення: 02.06.2026).

9.Meta AI Research. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. URL: https://ai.meta.com/research/publications/llama-guard-llm-based-input-output-safeguard-for-human-ai-conversations/ (дата звернення: 02.06.2026).

10. Springer / ICSOC 2024. Exploring the Systematic Use of LLMs for Microservices Generation. URL: https://link.springer.com/chapter/10.1007/978-981-96-7238-7_10 (дата звернення: 02.06.2026).

Цимбал Сергій Олегович – студент групи 4ПІ-24б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, м. Вінниця, e-mail: cimbal859@gmail.com

Бабюк Наталя Петрівна – доцент кафедри програмного забезпечення Вінницького національного технічного університету, м. Вінниця, e-mail: babiuk@vntu.edu.ua

Tsymbal Serhii O. – Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, e-mail: cimbal859@gmail.com

Babiuk Nataliia P. – Associate Professor of the Department of Software, Vinnytsia National Technical University, Vinnytsia, e-mail: babiuk@vntu.edu.ua