

ЕКОНОМІЧНІ АСПЕКТИ ВИКОРИСТАННЯ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ В БІЗНЕСІ

Сьогодні бізнес дедалі раціональніше підходить до використання моделей штучного інтелекту (ШІ-моделей). Якщо раніше компанії переважно експериментували з великими мовними моделями, оцінюючи їхній потенціал у розв'язанні різних бізнес-завдань, то нині ключового значення набуває економічна доцільність їх застосування.

Вартість використання великих мовних моделей в бізнесі можна оцінити через витрати на токени. Токен – це фрагмент послідовності символів, що може відповідати одній літері, частині слова, або цілому слову. Для одного й того самого тексту кількість tokenів в різних моделях може відрізнятися. Зазвичай виокремлюють вхідні токени, що відповідають сформульованому користувачем запиту, та вихідні токени, що відповідають згенерованій моделлю відповіді. Кожен розробник ШІ-моделей визначає власні тарифи на використання моделей, з якими можна ознайомитися на відповідних сайтах, зокрема [1-3].

Наприклад, для моделі gpt-5.5 компанії OpenAI у стандартному режимі використання API (Application Programming Interface, інтерфейс програмування застосунків) та короткого контексту, тобто за відносно невеликого обсягу вхідної інформації в одному запиті, вартість 1 млн вхідних tokenів становить 5 дол. США, а 1 млн вихідних tokenів – 30 дол. США [1]. Для моделі Claude Opus 4.8 компанії Anthropic у стандартному режимі вартість 1 млн вхідних tokenів становить 5 дол. США, а 1 млн вихідних tokenів – 25 дол. США [2]. Для моделі gemini-3.5-flash компанії Google у стандартному режимі вартість 1 млн вхідних tokenів становить 1,5 дол. США, а 1 млн вихідних tokenів – 9 дол. США [3].

Окрім того, існує відмінність і між тарифами на використання різних моделей одного розробника. Наприклад, для моделі gpt-5.5-pro у стандартному режимі короткого контексту 1 млн вхідних tokenів коштує 30 дол. США, а 1 млн вихідних tokenів - 180 дол. США. Натомість для моделі gpt-5.4-mini відповідно тарифи становлять 0,75 дол. США та 4,5 дол. США, а для gpt-5.4-nano – 0,20 дол. США та 1,25 дол. США. Така різниця у вартості використання моделей одного розробника пояснюється їх технічними відмінностями, зокрема, швидкістю виконання запиту, розміром контекстного вікна, максимальною довжиною відповіді, здатністю виконувати завдання різного рівня складності.

Серед зазначених моделей gpt-5.5-pro є моделлю з високим рівнем міркування та великим розміром контекстного вікна. Її доцільно використовувати для складних задач, однак вона потребує значних обчислювальних ресурсів і в окремих випадках може формувати відповідь протягом кількох хвилин [4]. Натомість gpt-5.4-mini та gpt-5.4-nano мають нижчий рівень міркування та менший розмір контекстного вікна, але забезпечують швидші відповіді [5, 6].

Сьогодні розробники пропонують значну кількість ШІ-моделей, а тому перед фахівцями компаній, які впроваджують технології штучного інтелекту, постає насамперед проблема обґрунтованого вибору ШІ-моделі. В кожному конкретному випадку такий вибір є складним, зокрема, через те, що ШІ-технології дуже швидко розвиваються і актуальність деяких моделей і побудованих за їх допомогою рішень не тривала. Тому компаніям необхідно щоразу шукати баланс між витратами на використання ШІ-моделі, якістю та швидкістю її відповідей, а також обсягом даних, які вона здатна опрацювати.

Для щоденного опрацювання масових бізнес-операцій доцільно використовувати дешевші й швидші ШІ-моделі за умови забезпечення достатньої якості відповідей, тоді як для складних управлінських рішень варто залучати дорожчі й потужніші моделі. Наприклад, для відповіді на типові запитання клієнтів щодо характеристик товару, умов доставки, статусу замовлення, для визначення категорії клієнтського звернення, для витягування з документів структурованих

даних цілком може бути достатньо простішої й дешевшої ШІ-моделі. У таких випадках ключове значення мають швидкість відповіді, низька вартість одного запиту та достатній рівень якості результату.

Окрім того, хорошою практикою є використання не однієї єдиної ШІ-моделі для усіх завдань компанії, а підбір різних ШІ-моделей до різних завдань відповідно до їх (завдань) складності, вартості помилки для бізнесу, вимог до швидкості, рівня масштабованості тощо. Особливо актуальним цей підхід стає в контексті побудови та використання ШІ-агентів, в яких ШІ-моделі можуть залучатися на кожному етапі виконання завдання. У такому разі використання дешевших моделей для простих операцій і дорожчих моделей для складніших завдань є економічно обґрунтованим. Зокрема, у роботі [7] представлено підхід до оптимізації використання великих мовних моделей, за якого запити під час виконання динамічно спрямовуються до дешевшої або потужнішої моделі залежно від складності задачі, що дає змогу істотно знизити витрати без суттєвої втрати якості відповідей.

Водночас потрібно розрізняти поняття «велика мовна модель» (BMM, Large Language Model, LLM), «мала мовна модель» (MMM, Small Language Model, SLM) та «мала версія великої мовної моделі». Значна частина моделей, доступних сьогодні на ринку, є великими мовними моделями або їхніми меншими версіями. Основна відмінність полягає в тому, що великі мовні моделі та їх менші версії зазвичай розміщені на серверах їх розробників, а обробка запитів користувачів відбувається за допомогою хмарних технологій, а іншим розробникам такі моделі переважно доступні через API. Натомість малі мовні моделі можуть бути локально розгорнуті на периферійних пристроях з обмеженими ресурсами, наприклад, на смартфоні, ноутбукі, або персональному комп'ютері, завдяки чому обробка запитів користувачів може відбуватися безпосередньо на такому пристрої.

Використання малих мовних моделей у ШІ-агентах та агентом ШІ може бути економічно й архітектурно доцільнішим, оскільки в агентних системах значна частина запитів до ШІ-моделі є простими й повторюваними [8]. У дослідженні [9] показано, що малі мовні моделі завдяки швидкій обробці запитів, нижчій затримці та меншому енергоспоживанню є придатними для використання на периферійних пристроях з обмеженими ресурсами. Водночас їхніми основними обмеженнями порівняно з великими мовними моделями є менша глибина міркування, слабше контекстне розуміння та вужче охоплення знань [9].

Отже, економічне обґрунтування використання ШІ-моделей в бізнесі полягає не лише у виборі найдешевшої моделі, а у визначенні оптимального співвідношення між вартістю, якістю, швидкістю, масштабованістю та ризиком помилки. Дешевші й менші моделі можуть бути ефективними для масових і повторюваних бізнес-операцій, тоді як потужніші і дорожчі моделі доцільно використовувати для складних аналітичних, управлінських або високоризикових задач. Тому перспективним підходом є побудова багаторівневої системи використання ШІ-моделей, у якій кожна модель застосовується відповідно до складності й економічної значущості конкретного завдання.

Список використаних джерел

1. OpenAI. (n.d.). Pricing. OpenAI Developers. URL: <https://developers.openai.com/api/docs/pricing>
2. Anthropic. (n.d.). Pricing. Claude API Docs. URL: <https://platform.claude.com/docs/en/about-claude/pricing>
3. Google. (n.d.). Gemini Developer API pricing. Google AI for Developers. URL: <https://ai.google.dev/gemini-api/docs/pricing>
4. OpenAI. (n.d.). GPT-5.5 pro. URL: <https://developers.openai.com/api/docs/models/gpt-5.5-pro>
5. OpenAI. (n.d.). GPT-5.4 mini. URL: <https://developers.openai.com/api/docs/models/gpt-5.4-mini>
6. OpenAI. (n.d.). GPT-5.4 nano. URL: <https://developers.openai.com/api/docs/models/gpt-5.4-nano>
7. Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2025). RouteLLM: Learning to Route LLMs with Preference Data. ICLR 2025. URL: <https://arxiv.org/pdf/2406.18665>

8. Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., & Molchanov, P. (2025). Small Language Models are the Future of Agentic AI. arXiv. URL: <https://arxiv.org/html/2506.02153v1>

9. Mahmoud, A. R. (2026). Small Language Models (SLMs) vs. LLMs: Efficiency and Accuracy on Edge Devices. ResearchGate. DOI: <https://doi.org/10.13140/RG.2.2.18176.70402>

Скіцько В.І.,

к.е.н., доцент,

Невінчана Д.О.,

здобувач першого (бакалаврського) рівня вищої освіти,

Київський національний економічний

університет імені Вадима Гетьмана

ЧАТ-БОТИ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ ЯК ЗАСІБ ВЗАЄМОДІЇ З КЛІЄНТАМИ В ЕЛЕКТРОННІЙ КОМЕРЦІЇ

У сфері електронної комерції зростання масштабів онлайн-торгівлі посилило потребу у використанні штучного інтелекту та аналітики великих даних для прогнозування поведінки споживачів, персоналізації рекомендацій, покращення клієнтського досвіду та підвищення ефективності продажів [1].

Одним із ключових інструментів сьогодишньої взаємодії з клієнтами в електронній комерції є чат-боти на основі штучного інтелекту (або ШІ-чат-боти) – програми, що імітують людське спілкування та, на відміну від традиційних сценарних чат-ботів, здатні розуміти запити користувачів і формувати автоматизовані відповіді на основі аналізу текстових даних [3]. У сучасних умовах поняття «ШІ-чат-боти» використовується у розширеному значенні та охоплює не лише базові діалогові системи, а й інтелектуальні рішення з елементами персоналізації, рекомендаційних функцій та автоматизованого підбору товарів.

За даними Juniper Research, завдяки появі розмовного та генеративного штучного інтелекту, обсяг світових роздрібних продажів, здійснених за допомогою чат-ботів (у тому числі ШІ-чат-ботів), зросте з 12 млрд дол. США у 2023 році до 72 млрд дол. США у 2028 році [2]. Така динаміка свідчить про поступове перетворення ШІ-чат-ботів з допоміжного інструменту на повноцінний елемент цифрової стратегії бізнесу.

За аналогією до класифікації чат-ботів, запропонованої для фінансового сектору, у сфері електронної комерції ШІ-чат-боти доцільно класифікувати за функціональним призначенням на інформаційні, транзакційні, консультаційні, лідогенеруючі та боти зворотного зв'язку. Вони відповідно забезпечують надання довідкової підтримки, автоматизацію типових операцій, персоналізацію взаємодії, збір даних потенційних клієнтів і отримання відгуків для покращення сервісу [3].

Технологічною основою сучасних ШІ-чат-ботів є обробка природної мови (Natural Language Processing, NLP), машинне навчання (Machine Learning, ML) та великі мовні моделі (Large Language Models, LLM). Вони забезпечують інтерпретацію й оперативну обробку запитів клієнтів, аналіз поведінкових даних користувачів та формування індивідуальних рекомендацій, що сприяє підвищенню якості обслуговування, утриманню клієнтів і підвищує комерційну ефективність бізнесу [4].

Використання ШІ-чат-ботів поступово набуває поширення в діяльності провідних компаній. Наприклад, у 2026 році Starbucks запустила бета-версію розмовного застосунку (чат-бота у розширеному функціональному розумінні) у середовищі ChatGPT, який рекомендує напої на основі настрою або фото клієнта і дозволяє модифікувати замовлення безпосередньо в чаті. Як зазначає компанія на своїй офіційній веб-сторінці, цей інструмент – це “технологія, яка розкриває творчі здібності та допомагає клієнтам відкривати щось