**УДК: 004.8** https://doi.org/10.15407/jai2025.03.032

**А. О. Никоненко**
Черкаський державний технологічний університет, Україна
б-р Шевченка, 460, м. Черкаси, 18006
andrey.nikonenko@gmail.com
https://orcid.org/0000-0002-9442-1601

# СИСТЕМАТИЧНИЙ ОГЛЯД ДОСЯГНЕНЬ В ГАЛУЗІ LLM ВІД GPT-3 ДО МІРКУЮЧИХ АГЕНТІВ

**Анотація.** Стаття пропонує комплексний аналіз етапів еволюції великих мовних моделей у період з 2020 по 2025 рік. Ми проаналізували стратегічний фокус, підходи, архітектуру та ключові етапи розвитку індустрії ШІ з фокусом на внеску десяти провідних компаній: OpenAI, Anthropic, Google, Meta, Mistral AI, DeepSeek AI, AI21 Labs, Qwen, Cohere та xAI. Проаналізовано основні парадигми розвитку LLM: масштабування, запропоновану з виходом моделі GPT-3; узгодження відповідей моделі із запитами користувача, що характеризувалася появою технік на кшталт RLHF; поширення моделей з відкритими вагами; зниження вартості моделей шляхом оптимізації; та сучасний етап, орієнтований на ефективність архітектур, мультимодальність і агентну поведінку. Визначено ключові тенденції, що сформували поточний ландшафт генеративного ШІ, включаючи конкуренцію між пропрієтарними та відкритими моделями, перехід від універсальних моделей до спеціалізованих агентів та зростання важливості економічної ефективності. Робота узагальнює поточний стан індустрії за допомогою ієрархічної моделі рівнів розвитку технологій та окреслює перспективи подальшого розвитку галузі, підкреслюючи роль агентних екосистем та якісних даних.

**Ключові слова:** великі мовні моделі, генеративний ШІ, еволюція ШІ, архітектура моделей, агентне мислення, огляд.

**A. Nykonenko**
Cherkasy State Technological University, Ukraine
Shevchenka Blvd, 460, Cherkasy, 18000
andrey.nikonenko@gmail.com
https://orcid.org/0000-0002-9442-1601

# A SYSTEMATIC REVIEW OF LLM ADVANCEMENTS FROM GPT-3 TO REASONING AGENTS

**Annotation.** The article offers a comprehensive analysis of the stages of evolution of large language models between 2020 and 2025. We analyzed the strategic focus, approaches, architecture, and key stages of development of the AI industry, focusing on the contributions of ten leading companies: OpenAI, Anthropic, Google, Meta, Mistral AI, DeepSeek AI, AI21 Labs, Qwen, Cohere, and xAI. We analyzed the main paradigms of LLM development: scaling, proposed with the release of the GPT-3 model; aligning model responses with user prompts, characterised by the emergence of techniques such as RLHF; spreading models with open weights; reducing model costs through optimisation; and the current stage, focused on architectural efficiency, multimodality, and agent behaviour. Key trends that have shaped the current landscape of generative AI are identified, including competition between proprietary and open models, the rise of specialised agents over universal models, and the growing importance of economic efficiency. The paper summarises the current state of the industry on a proposed hierarchical model of technology development levels and outlines prospects for further development of the industry, emphasising the role of agent ecosystems and high-quality data.

**Keywords:** large language models, generative AI, artificial intelligence, AI evolution, model architecture, agentic reasoning, review.

## Постановка проблеми

Поява великих мовних моделей (LLM) на початку 2020-х років являє собою значну подію в історії розвитку штучного інтелекту. Можна назвати цей період «новим літом» ШІ, такі періоди характеризуються появою нових інтелектуальних технологій, які привертають велику увагу до галузі та ведуть до притоку інвестицій. Інвестиції, в свою чергу, ведуть до ще більш активного

розвитку технологій, підвищення їх практичної цінності та проникнення в різні сфери життя людини [1]. Генеративні моделі всього за кілька років пройшли шлях від лабораторних експериментів до практичних застосунків з аудиторією в десятки і сотні мільйонів людей. За ці п'ять років було винайдено, практично використано і визнано застарілими різні архітектурні концепції: створення осмислених текстів, масштабування розміру моделей, побудова агентів-співрозмовників, фокус на фактах і достовірності, ефективності та безпеці. З плином часу технологія завойовує все нові ринки, а конкуренція між виробниками фундаментальних моделей тільки наростає. Експерти попереджають про ризики і вимагають більш жорсткої регуляції галузі. Високі темпи розвитку ШІ разом з його широким впровадженням, часто без достатнього аналізу надійності, вимагають виваженого та системного аналізу галузі для розуміння її поточних можливостей, траєкторії руху та потенційних викликів.

**Мета**

Метою статті є систематизація наявних підходів та парадигм побудови генеративних мовних моделей, створених протягом останніх 5-ти років, на прикладі десяти провідних гравців галузі. Дослідження спрямоване на аналіз та порівняння стратегічного фокусу різних компаній, їх вибір архітектури, ключових принципів та способів застосування моделей. Також, однією з цілей даного дослідження є спроба розібратися, чи існує єдиний мейнстрімний підхід до побудови LLM, чи кожен гравець діє в рамках свого унікального напрямку. Іншою ціллю є визначення основних технологічних парадигм побудови великих мовних моделей та векторів майбутніх досліджень.

**Методи**

Основним способом дослідження є аналітичний огляд з використанням методів систематичного та порівняльного аналізу. Перераховані методи застосовуються до **об'єкта дослідження** - десяти компаній-розробників фундаментальних моделей: OpenAI, Anthropic, Google, Meta AI, Mistral AI, xAI, DeepSeek, Qwen, AI21 Labs та Cohere. Таким чином ми вивчаємо **предмет дослідження** - технології та стратегії побудови LLM. Порівняльний аналіз використовується для співставлення стратегій позиціювання на ринку, вибору архітектури моделей та цілей гравців. Шляхом синтезу отриманої інформації було побудовано ієрархічну модель рівнів розвитку ШІ та виділено основні тенденції, що будуть направляти розвиток галузі в найближчі роки.

**OpenAI**

Базу для сучасного швидкого розвитку технологій генеративного ШІ було закладено компанією OpenAI. Їх дослідження, присвячені побудові архітектури генеративних попередньо навчених трансформерів (GPT), можна вважати фундаментом, на якому стоїть вся сучасна індустрія генеративного ШІ. Статті "Improving Language Understanding by Generative Pre-Training" [2] та "Language Models are Unsupervised Multitask Learners" [3] є ключовими роботами OpenAI, що заклали цю базу. Перша стаття присвячена випуску моделі GPT-1; в ній запропонована концепція загальноцільової моделі (task-agnostic model). Основна особливість даної моделі - можливість навчання на наборі нерозмічених даних, з подальшим використанням невеликої кількості розмічених даних для виконання специфічних завдань. Другу статтю присвячено GPT-2, в ній показано, що збільшення розміру моделі та кількості навчальних даних відкриває можливості для адаптації до конкретних завдань без попереднього навчання (zero-shot task transfer). Ці два дослідження стали ключовими при створенні генеративного ШІ.

Першим кроком до широкого успіху та визнання компанії OpenAI стала «гіпотеза масштабування», яку було запропоновано в

статті "Language Models are Few-Shot Learners" [4]. Успіх GPT-3, описаний в статті, закріпив масштабування як головний напрям досліджень. Також даний реліз досяг нового рівня якості, встановивши бенчмарк для майбутніх моделей. Крім суттєвого успіху були і певні проблеми, наприклад, непередбачуваність виходів моделі та генерація шкідливого контенту. Для подолання цих проблем було створено механізм узгодження між запитами користувача та виходами моделі, що було запропоновано в роботах "Summarizing Books with Human Feedback" [5] та "Training language models to follow instructions with human feedback" [6]. В цих статтях було запропоновано метод RLHF, який було використано при тренуванні моделі InstructGPT.

Наступними кроками до покращення можливостей моделей стало запровадження здатності до багатокрокового міркування та поява мультимодальності. Мультимодальність було презентовано в технічному звіті [7], що було присвячено виходу GPT-4 у березні 2023 року. Нова модель була здатна однаково якісно працювати з вхідними даними як у вигляді тексту, так і у вигляді зображення. Продовжуючи тему дослідження можливостей з міркування та сприйняття світу моделями, компанія створила модель Sora, представлену в лютому 2024 року. Sora була здатна перетворювати текст на відео і позиціонувалася як симулятор світу. У вересні 2024 року було представлено модель o1 [8], а у квітні 2025 року моделі o3 та o4-mini [9]. Ця серія LLM використовує глибокі міркування (reasoning) як окрему спеціальну здібність.

Послідовність в дослідженнях та розвиток попередніх результатів забезпечили компанії лідерство в індустрії. Реліз GPT-1 заклав базу для подальших досліджень в сфері генеративного ШІ, хоча і не мав особливої практичної цінності. GPT-2 став наступним кроком і продемонстрував можливості моделей з адаптації до нових завдань без донавчання. GPT-3 базувався на раніше отриманих результатах, був більш придатним до практичного використання і показав, що масштабування моделей веде до підвищення їх якості. InstructGPT продемонстрував, що використання технік підлаштування моделей під очікування користувача значно покращує їх безпечність, що веде до росту популярності ШІ-агентів серед користувачів. Реліз GPT-4 став наступним кроком розвитку генеративного ШІ, представивши мультимодальні моделі і моделі серії «o». Останніх п’ять років OpenAI є безумовним лідером у дослідженнях ШІ і встановлює стандарти індустрії.

**Anthropic**

Основу колективу Anthropic склали люди, що пішли з OpenAI у 2021 році. Причини розбіжностей між командою Anthropic та їх колегами з OpenAI достеменно невідомі, однак можна припустити, що основною підставою відокремлення було питання безпеки штучного інтелекту. Відомо, що ще до відокремлення були дискусії щодо ризиків комерціалізації моделей і неконтрольованого розширення можливостей ШІ без проведення детальних перевірок. Додатковим підтвердженням може слугувати той факт, що CEO Anthropic підписав відкрий лист «Pause Giant AI Experiments: An Open Letter». В Anthropic наголошують, що подальший розвиток штучного інтелекту не може базуватися на сприйнятті LLM як чорного ящика з подальшим контролем його результатів. Створення безпечних моделей вимагає глибокого розуміння процесів, що відбуваються всередині нейронної мережі.

Початкову фазу досліджень було присвячено створенню аналітичних засобів для вивчення особливостей роботи трансформерів. У статті "A Mathematical Framework for Transformer Circuits" [10] було запропоновано математичний апарат, що дозволяє відтворити операції, що відбуваються в трансформерах, та введено поняття "circuits" та "induction heads". Продовження роботи над цією тематикою було представлено у статті "Toy Models of

Superposition" [11], де було визначено, яким чином нейронні мережі можуть мати кількість фіч більшу за кількість нейронів.

Наступний етап розвитку характеризується застосуванням принципу "безпека понад усе" до практики тренування моделей. В статті "Claude's Constitution" [12] було запропоновано метод навчання нешкідливого ШІ без застосування людей-розмітчиків, базуючись на самокритиці моделі на основі невеликого статуту (constitution) основних принципів. Наразі компанія зосереджена на створенні та поширенні стандартів безпеки в індустрії та використанні раніше розроблених засобів інтерпретованості для аналізу сучасних моделей. У статті "Tracing the thoughts of a large language model" [13] показано застосування технік інтерпретованості до моделі Claude 3.5 Haiku. У листопаді 2024 року компанія представила відкритий стандарт "Model Context Protocol", що дозволяє уніфікувати підключення зовнішніх джерел даних. Крім того, стаття "RAIL in the Wild" [14] демонструє використання незалежними експертами використання відкритого набору даних Anthropic для оцінки етичності поведінки моделей.

На шляху свого розвитку компанія продовжує дотримуватися ключового принципу, що було проголошено при її створенні: безпечний ШІ можна створити лише через глибоке і точне розуміння процесів, що відбуваються всередині моделі. Фокус на інтерпретованості дає надію на створення прозорих, надійних та прогнозованих систем генеративного ШІ.

**Google**

Компанія Google займається дослідженнями ШІ вже досить довгий термін, на її рахунку такі відомі проєкти, як AlphaGo, TensorFlow, AlphaFold 3 та багато інших. Мабуть, компанія має один з найширших профілів наукових інтересів і один з найбільших внесків, пов’язаних зі ШІ. Незважаючи на велику кількість досліджень, можна стверджувати, що Google запізно вступив до конкурентної боротьби в царині генеративного ШІ і програв перший етап меншим і більш молодим конкурентам.

Незважаючи на велику кількість проєктів, пов’язаних з NLP, наприклад, BERT [15] і T5 [16], перші експерименти з LLM були не дуже успішними. Іноді це призводило навіть до курйозів, так, наприклад, реліз ШІ-агента BARD [17], що працював на базі моделі LaMDA [18], супроводжувався численними галюцинаціями, деякі з яких було зафіксовано навіть під час презентації. Модель PaLM [19], що представляла альтернативний напрямок досліджень, дозволила компанії створити прямого конкурента GPT-3 але не перевершити його. Випуск PaLM 2 [20] символізував наступну спробу наздогнати конкурентів. На жаль, реліз цієї моделі, здатної працювати виключно з текстовими даними, відбувся через кілька місяців після виходу мультимодальної GPT-4, що лише акцентувало суттєве відставання Google від state-of-the-art на той момент.

Google переглянув стратегію ШІ і консолідував зусилля різних дослідницьких груп, в результаті чого вийшла стаття "Gemini: A Family of Highly Capable Multimodal Models" [21]. В цій роботі описана серія мультимодальних моделей Gemini (Ultra, Pro та Nano), здатних працювати з текстом, кодом, зображеннями, аудіо та відео. Результати тестів показали, що дана серія моделей вже здатна конкурувати з основними моделями свого часу.

Отже, незважаючи на великий вплив компанії на розвиток ШІ, Google недооцінила значущість створення перших LLM. Консолідувавши значні ресурси, компанія змогла наздогнати індустрію генеративного ШІ і зараз займає одну з провідних позицій. Google намагається утримувати широкий фокус, активно розвиваючи лінійку пропрієтарних моделей (останній реліз — Gemini 2.5) та вступивши в конкуренцію на ринку open source моделей з релізом серії моделей Gemma.

**Meta AI**

Компанія Meta зробила значний внесок в індустрію ШІ, включаючи такі широко відомі речі, як фреймворк PyTorch та бібліотеку FastText. Як і Google, Meta дещо затрималася з виходом своєї першої LLM. LLaMA [22] вийшла лише в лютому 2023 року. Витік вагових коефіцієнтів моделі зробив її першою справді відкритою LLM, хоча й випадково. Реліз "Llama 2: Open Foundation and Fine-Tuned Chat Models" [23] був першим свідомим open source релізом LLM від Meta. Модель програвала конкурентам на кшталт GPT-4, проте можливість вільного використання такої потужної моделі дала імпульс для появи великої кількості нових продуктів на базі ШІ.

Водночас компанія проводила дослідження в інших напрямках. Серед основних результатів паралельного треку можна виділити модель SAM для сегментації зображень, Toolformer для роботи з зовнішніми інструментами і SeamlessM4T в якості засобу для мультимодального перекладу. Meta швидко стала лідером в галузі open-source LLM і зосередилася на якості моделей. Нове покоління моделей, LLaMA 3 [24] було представлено в 2024 році. В реліз входила велика кількість моделей, найбільша з яких мала 405B параметрів і була здатна конкурувати за якістю з GPT-4. У поколінні LLaMA 3 було кілька інкрементальних релізів. Сімейство LLaMA 3.1 [25] містило модель на 405B параметрів, сімейство LLaMA 3.2 [26] містило мультимодальні моделі, а LLaMA 3.3 [27] містило покращені механізми розуміння та слідування інструкціям.

Отже, спочатку випадковість, а потім і стратегічне рішення сфокусуватися на open source дозволили Meta зайняти одну з провідних позицій на ринку. Компанія продовжує розвивати свої моделі в напрямку мультимодальності та у використанні нових архітектур, наприклад, нещодавній реліз серії LLaMA 4 [28] базується на архітектурі MoE. Інший важливий напрямок - ускладнення поведінки моделей шляхом переходу від простого слідування інструкціям до багатокрокового міркування та агентної поведінки.

**Фокус на ефективності: Mistral AI, DeepSeek AI та AI21 Labs**

Вже за кілька років почали вироблятися стандарти в галузі генеративного ШІ. Провідні гравці переймали успішні техніки та підходи один в одного і зрештою створення LLM перетворилося на набір стандартних методів, вироблених компаніями, що прийшли в індустрію першими. Стандартизація підходів не лишила місця новим, меншим гравцям, які не могли змагатися з гігантами в парадигмі масштабування, тож єдиним альтернативним шляхом для них було сфокусуватися на ефективності. Нові гравці не фокусувалися на потраплянні в топ лідерборду, проте вони намагалися запропонувати моделі, що демонструють прийнятну якість за помірні кошти. Ключем до такого переходу стала оптимізація всіх процесів: вибору архітектури, тренування моделі та сценаріїв використання.

**Mistral AI**

Mistral AI приєднався до гонки генеративних моделей в жовтні 2023 з виходом своєї першої моделі Mistral 7B [29]. Цю модель було випущено в open source. Завдяки застосованим методам оптимізації, таким як Grouped-Query Attention (GQA) і Sliding Window Attention (SWA), модель, незважаючи на досить невеликий розмір, змогла показати результати, кращі за LLaMA 2 13B та перемогти LLaMA 1 34B в деяких тестах. В тій же статті було представлено і модель Mistral 7B-Instruct, яка перевершила LLaMA 1 34B в більшості тестів. В наступні кілька років компанія випустила цілу серію моделей:

- пропрієтарні, наприклад, Mistral Large 2 [30] та Mistral Medium 3 [31];
- open source, наприклад, Mixtral 8x22B на базі архітектури MoE.

Створивши набір досить потужних базових моделей, компанія перейшла до наступного рівня, прагнучи отримати

мультимодальність та здатність до просунутого міркування. Прикладом мультимодальної моделі виступає Mistral OCR, що здатна розпізнавати документи, а роль моделі, здатної до міркування, виконує Magistral [32].

**DeepSeek AI**

Дослідження компанії DeepSeek AI тісно пов'язані з побудовою моделей, оптимізованих під HPC (high-performance computing) інфраструктуру. Основні результати викладено у серії публікацій "DeepSeek LLM" [33], "DeepSeek-V2" [34] та "DeepSeek-V3" [35]. Успіх моделей цієї серії було частково досягнуто за рахунок оптимізації наявних архітектур та використання даних, згенерованих іншими моделями, у якості навчальних, включаючи пропрієтарні [36].

DeepSeek LLM 67B перевершив за тестами LLaMA 2 70B. DeepSeek-V2 з 236B параметрів, побудована на архітектурі MoE, демонструвала якість, близьку до LLaMA 3 70B, використовуючи втричі менше параметрів під час інференсу. DeepSeek-V3 з 671B параметрів перевищила за більшістю показників таких лідерів, як Llama-3.1-405B та GPT-4o.

Таблиця 1. Порівняння цін флагманских моделей (січень 2025)

| Назва моделі | Розробник | Ціна за 1M вхідних токенів | Ціна за 1M вихідних токенів |
|---|---|---|---|
| DeepSeek-R1 | DeepSeek | $0.55 | $2.19 |
| Mistral Medium | Mistral AI | $2.70 | $8.10 |
| Mistral Large | Mistral AI | $4.00 | $12.00 |
| Jamba 1.5 Large | AI21 Labs | $2.00 | $8.00 |
| Jurassic-2 Mid | AI21 Labs | $12.50 | $12.50 |
| Jurassic-2 Ultra | AI21 Labs | $18.80 | $18.80 |
| OpenAI o1 | OpenAI | $15.00 | $60.00 |
| GPT-4o | OpenAI | $5.00 | $20.00 |
| GPT-4 (older versions) | OpenAI | $30.00 | $60.00 |
| Claude 3 Opus | Anthropic | $15.00 | $75.00 |
| Claude 3 Sonnet | Anthropic | $3.00 | $15.00 |
| Gemini 1.0 Pro | Google | $2.00 | $6.00 |

Мабуть, найбільш новаторською стала робота "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning" [37], в якій компанія представила свою першу модель, здатну до багатокрокових міркувань. Стаття описує, як DeepSeek використовує RL у великих масштабах для отримання можливостей покрокових міркувань напряму від базової моделі, без етапу supervised fine-tuning (SFT). Реліз цієї моделі в open source перевернув всю індустрію, змінивши наявну економічну модель. Вартість роботи DeepSeek-R1 була приблизно в 30 разів нижчою за OpenAI-o1, що поставило у складне положення більшість великих гравців. Детальне порівняння цін

моделей наведено в таблиці 1. Вихід моделі DeepSeek-R1 поклало початок суттєвому зниженню цін на ринку генеративного ШІ.

**AI21 Labs**

Компанія AI21 Labs добре відома завдяки своїм дослідженням LLM, що проводилися ще до загального буму. Серед ранніх моделей найбільш відома серія Jurassic-1, що конкурувала з GPT-3. Інший суттєвий внесок — одна з перших робіт щодо створення зовнішніх інструментів для LLM [38].

Сучасний внесок компанії найкраще представлений серією моделей Jamba. У статті "Jamba: A Hybrid Transformer-Mamba Language Model" [39] було запропоновано використовувати гібридну архітектуру, що складається з Transformer, Mamba та MoE. Архітектура Mamba є фундаментальною альтернативою Transformer, побудованою на базі Selective State Space Models (SSMs). Ключовими перевагами гібридної архітектури Jamba стала можливість забезпечити результати на рівні інших передових моделей при підтримці збільшеного до 256K токенів контекстного вікна та суттєво більша пропускна здатність. Наступні роботи "Jamba-1.5" та Jamba-1.6 представили подальші покращення та нові техніки оптимізації. Важливою особливістю серії Jamba стало те, що їх ваги було випущено в open source.

**Фокус на комерційному використанні: Qwen, Cohere та xAI**

Конкуренція за лідерство в таблиці лідерів за рахунок розміру моделі та обсягу обчислювальних ресурсів не гарантує комерційного успіху компанії-виробнику LLM. Питання практичної застосовуваності моделі виходить на передній план. Яку практичну користь або додану вартість модель може принести бізнесу? Чому для вирішення певної задачі користувач має обрати певну LLM? Всі три компанії, описані в даному розділі, об'єднані стратегією створення комер-ційних моделей для конкретних потреб бізнесу. Це дозволяє їм фокусуватися на певних нішах і відвойовувати свою долю ринку без прямої конкуренції з чотирма великими компаніями.

**Qwen**

Траєкторію досліджень компанії Alibaba можна розділити на дві стадії: створення сімейства моделей Qwen та архітектурні інновації з подальшим масштабуванням. Поточна стратегія компанії полягає у створенні якісного та доступного набору open source моделей з легкою інтеграцією, що покривають всі популярні потреби. Alibaba/Qwen намагається зайняти нішу єдиного уніфікованого провайдера відкритих моделей на всі випадки життя.

Початок досліджень було покладено в статті "Unifying Architectures, Tasks, and Modalities..." [40], яка запропонувала фреймворк OFA. Формально, команда Qwen приєдналася до LLM ралі у 2023 році з виходом перших моделей Qwen та Qwen-Chat. Сімейство моделей включало версії від 1.8B до 72B параметрів, а також спеціалізовані моделі для коду та математики. Реліз серії моделей Qwen1.5 відбувся у лютому 2024 року, в цю серію входила перша модель з архітектурою MoE. Серія моделей Qwen2, зі значними покращеннями багатомовних можливос-тей, вийшла у липні 2024 року. Серія Qwen2.5 містила не тільки набір покращених відкритих моделей, але й кілька пропрієтарних LLM (наприклад, Qwen2.5-Turbo, Plus, Max), які показували результати на рівні GPT-4o. Реліз серії Qwen3 [41] у травні 2025 року представив моделі, здатні до багатокрокового міркування з підтримкою 119 мов, найбільша модель серії мала 235B параметрів.

**Cohere**

Компанію Cohere було засновано одним з авторів статті "Attention is All You Need" [42], що вперше описала архітектуру трансформерів. З самого початку компанія фокусувала зусилля на створенні ШІ для бізнесу. Безпечність, масштабованість та точність при виконанні реальних задач стали ключовими елементами її стратегії.

Технологічний фокус компанії зосереджено на RAG (Retrieval-Augmented Generation) та можливості багатокрокового використання зовнішніх інструментів.

RAG - це спеціальна техніка введення в модель додаткових знань без донавчання. Цей підхід дозволяє LLM отримувати доступ до нових знань або внутрішньої інформації компанії шляхом включення її в запит користувача. Техніка включає наступні етапи:

- перетворення зовнішніх даних у векторне представлення та збереження в базі знань;
- пошук інформації, релевантної до запиту користувача;
- доповнення запиту цією інформацією перед подачею до LLM.

Такий підхід підвищує точність відповідей та актуальність даних, а також дозволяє зменшити кількість галюци-націй.

Генеративні моделі компанії, наприклад, medium та xlarge, вперше з'явилися в 2021 році, проте суттєвим кроком уперед стала модель Command R+, випущена компанією в 2024 році. Модель мала 104B параметрів та була оптимізована під використання у бізнес-сценаріях і RAG. Наступним етапом став реліз моделі Command A з 111B параметрів, що підтримує 23 мови та має покращені можливості RAG.

Іншою важливою областю для компанії є розвиток багатомовної підтримки. Компанія веде відкритий дослідницький проєкт Aya. В 2024 році проєкт представив модель, здатну працювати зі 101 мовою, перевершивши інші подібні ініціативи, наприклад, BLOOMZ. Поточні дослідження компанії сфокусовані на багатокроковому міркуванні для різних мов та мультимодальності [43]. Також компанія працює над покращенням практик оцінювання LLM, що демонструє робота "The Leaderboard Illusion" [44].

**xAI**

Підхід компанії xAI, заснованої Ілоном Маском у 2023 році, суттєво відрізняється від інших. Тут основну ставку зроблено на швидкість розробки, тому багато аспектів, що є важливими для інших гравців (наприклад, безпека моделей), може бути принесено в жертву. Велика база пропрієтарних даних з платформи X та обчислювальних потужностей (суперкомп'ютер Colossus) дозволяє компанії робити часті релізи нових версій моделей.

Основною моделлю компанії є Grok. Перший реліз, Grok-1, відбувся в 2023 році, це модель на 314B параметрів на архітектурі MoE. Модель запам'яталася своєю незвичною особистістю, створеною на базі "The Hitchhiker's Guide to the Galaxy". Згодом ваги моделі було викладено в open source. Наступні версії, Grok-1.5 та Grok-1.5 Vision, додали покращені можливості мислення та мультимодальність. Grok-2 продемонструвала суттєве покращення продуктивності, а модель Grok-3 Beta отримала збільшене до 1 мільйона токенів контекстне вікно та режими для міркувань (Think) та досліджень (DeepSearch). Останній реліз, Grok-4, просувається як модель з міркуваннями на рівні PhD, вона заснована на MoE архітектурі з приблизно 1.7T параметрів, підтримує мульти-модальність в реальному часі та можливості multi-agent parallel reasoning.

Підхід xAI зазнає значної критики. Дослідження виявили високу вразливість до jailbreaking, маніпулятивну поведінку, створення шкідливого контенту та відсутність прозорості [45]. Незважаючи на присутню критику, компанія продовжує швидко рухатися вперед, забезпечивши собі місце серед лідерів індустрії.

**Висновки**

Період розквіту генеративного ШІ почався в 2020 році з виходом GPT-3, за кілька років він привернув значну увагу великого кола користувачів та інших компаній і поклав початок гонки ШІ. Певний час змагання йшло за принципом "чим більше, тим краще", але через залученням більшої кількості компаній, парадигма змінилася і відбувся розквіт різних

архітектур, оптимізацій та спеціалізованих моделей.

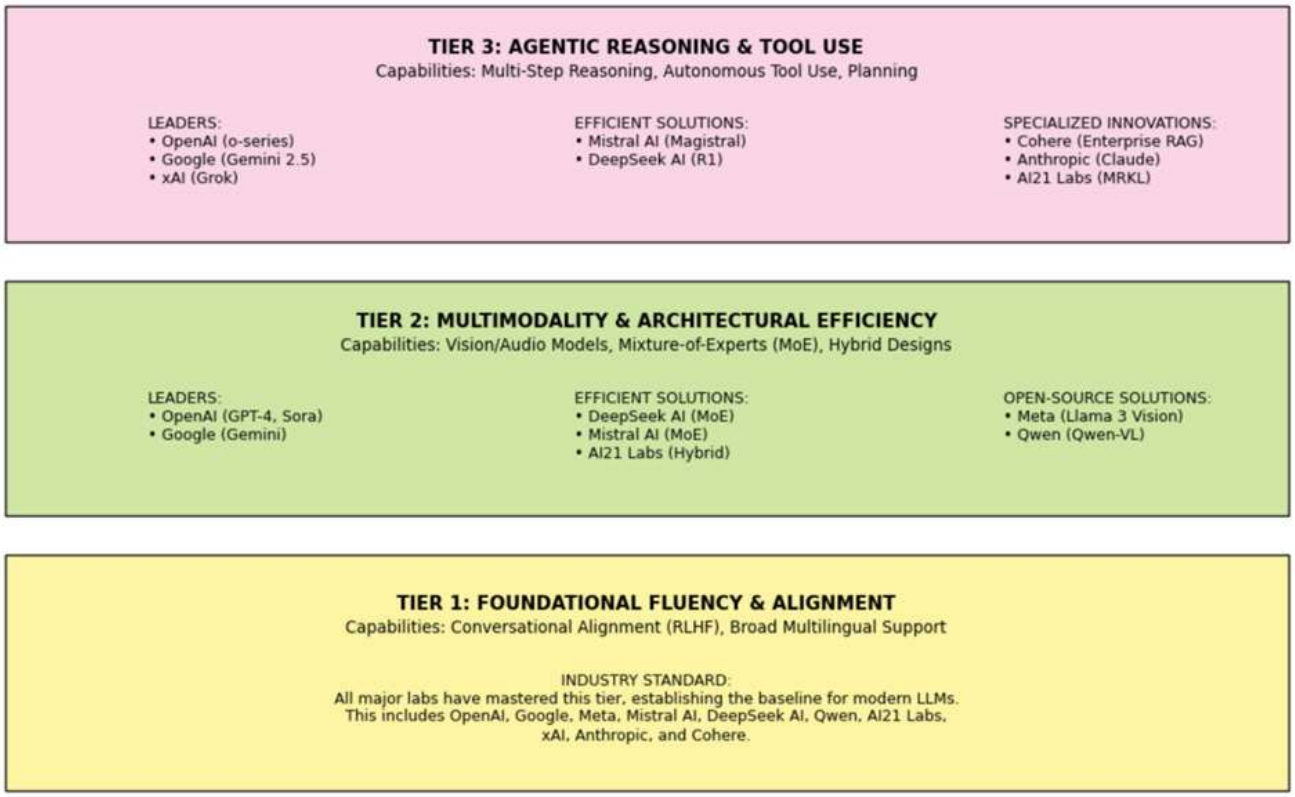


Рис. 1. Рівні розвитку технологій ШІ

Індустрія не фокусується на одному рішенні, натомість, кожен з основних гравців має свої цілі і спрямовує зусилля і дослідження для зайняття своєї ніші на ринку.

Якщо спробувати розділити поточні досягнення в світі ШІ на умовні рівні розвитку технології та спробувати розділити основних гравців за цією ієрархією, то ми отримаємо приблизно таку схему (рис.1). Варто відмітити, що навіть в середині одного рівня рішення різних компаній можуть суттєво відрізнятися, як результат стратегічних пріоритетів. Наприклад, на третьому рівні ми маємо 8 компаній, при цьому три з них сфокусовані на утриманні найвищих місць в таблиці лідерів за тестами, дві - на ефективності моделей, а решта пропонують спеціалізовані нішеві рішення. Проте, всі перелічені компанії об'єднує те, що вони мають моделі, здатні до багатокрокового мислення та використання зовнішніх інструментів. До того ж важко однозначно визначити, чи відноситься конкретна компанія до певного рівня, чи ні. Наприклад, xAI формально може бути на всіх рівнях, хоча зовнішні дослідження і показують суттєві проблеми з узгодженням між відповідями моделі та користувацьким запитам (tier 1), а остання модель компанії Grok-4 підтримує мультимодальність, проте не є MoE (tier 2). Тобто, можна включити xAI в tier 2 і виключити з tier 1. Незважаючи на певні недоліки, запропонована схема найкраще відтворює поточний прогрес в галузі генеративного ШІ.

Якщо говорити про ймовірні перспективи розвитку ШІ, то варто вказати кілька потенційних напрямків:

1. Поява агентних екосистем. Вже зараз наявні сценарії, де LLM виходить за рамки ролі чат-бота. В таких ситуаціях LLM починає грати роль "мозку", що керує набором інших інструментів та агентів для

планування і виконання складних, багатокрокових задач.

2. Подальша фрагментація ринку. Епоха однієї моделі-експерта, здатної однаково добре виконувати всі задачі, закінчується. Варто очікувати появи спеціалізованих моделей для досліджень, програмування, медицини, юриспруденції тощо.

3. Ефективність витрат. Моделі стають інструментами для виконання задач бізнесу, тому майбутнє за оптимізацією витрат та збільшенням ефективності і точності в конкретних задачах.

4. Дані, як ключовий фактор. Більшість передових архітектур сьогодні доступні в open source, а техніки покращення моделей описано в наукових публікаціях. В той самий час, практично всі дані, доступні онлайн, вже було використано для навчання. В такій ситуації саме доступ до пропрієтарних високоякісних даних та їх синтетичних аналогів стає ключовим фактором для побудови наступної генерації моделей.

Прогрес в галузі генеративного ШІ йде з величезною швидкістю. Адаптація нових архітектур і рішень, що раніше могла займати роки, тепер відбувається дуже швидко. Те, що створено одним гравцем, вже в наступній ітерації моделей буде застосовано всіма іншими. Незважаючи на це, архітектура трансформерів залишається лідируючим рішенням. Серед архітектурних підходів останньою інновацією являється широке впровадження MoE та експерименти з Transformer-Mamba підходом.

Аналізуючи питання можливої скорої потенційної появи AGI (Artificial General Intelligence), ми припускаємо, що цього не станеться. Безумовно, LLM — це надзвичайно потужна технологія, що вже змінила наше сприйняття інтелектуальних систем. Мовні моделі тільки на початку свого розвитку і за кілька років вони будуть значно потужнішими. Проте вже зараз ми розуміємо, що в цих моделей відсутнє справжнє розуміння і свідомість, тож вони не зможуть розвинутися в AGI. Генеративний ШІ — передова інтелектуальна технологія, що може призвести до скорої появи ШІ-лікарів, юристів, вчителів та персональних асистентів, здатних сприймати світ навколо вас в режимі реального часу. Вони можуть змінити на краще наш спосіб взаємодії зі світом, проте вони не мають справжнього розуму.

Еволюція моделей генеративного ШІ від GPT-1 до багатомовних, мультимодальних систем, здатних використовувати інструменти та проводити багатокрокове мислення, демонструє не лише суттєвий технологічний прогрес, але і змінює наше розуміння інтелекту як такого. Значного розвитку зазнала і галузь ШІ в цілому - те, що починалося як змагання за найвище місце в таблиці лідерів, отримало широке практичне застосування і переросло в новий великий ринок. Тепер тут існують різні напрямки і ніші, кожна з яких пропонує свій унікальний баланс потужності, безпеки, доступності та ефективності. Розвиток генеративного ШІ продовжується і ми очікуємо появи ще більш корисних та автономних моделей-помічників, здатних надавати кваліфіковану допомогу людям в певних галузях.

## Література (References)


1. Nykonenko, A. (2023). *The impact of artificial intelligence on modern education: prospects and challenges.* Artificial Intelligence, 2, 10-15.
2. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.
3. Radford, A., Wu, J., Child, R., et al. (2019). *Language models are unsupervised multitask learners*. OpenAI blog, 1(8), 9.
4. Brown, T., Mann, B., Ryder, N., et al. (2020). *Language models are few-shot learners.* Advances in neural information processing systems, 33, 1877-1901.
5. Wu, J., Ouyang, L., Ziegler, D. M., Stiennon, N., Lowe, R., Leike, J., & Christiano, P. (2021). *Recursively summarizing books with human feedback*. arXiv preprint arXiv:2109.10862.
6. Ouyang, L., Wu, J., Jiang, X., et al. (2022). *Training language models to follow instructions with human feedback*. Advances in neural information processing systems, 35, 27730-27744.
7. Achiam, J., Adler, S., Agarwal, S., et al. (2023). *Gpt-4 technical report*. arXiv preprint arXiv:2303.08774.

8. OpenAI. (2024). *Learning to reason with LLMs*. OpenAI. https://openai.com/index/learning-to-reason-with-llms/

9. OpenAI. (2025). *Introducing OpenAI o3 and o4-mini*. OpenAI. https://openai.com/index/introducing-o3-and-o4-mini/

10. Anthropic. (2021). *A Mathematical Framework for Transformer Circuits*. Anthropic. https://www.anthropic.com/research/a-mathematical-framework-for-transformer-circuits

11. Anthropic. (2022). *Toy Models of Superposition*. Anthropic. https://www.anthropic.com/research/toy-models-of-superposition

12. Anthropic. (2023). *Claude's Constitution*. Anthropic. https://www.anthropic.com/news/claudes-constitution

13. Anthropic. (2025). *Tracing the thoughts of a large language model*. Anthropic. https://www.anthropic.com/research/tracing-thoughts-language-model

14. Verma, S., Prasun, P., Jaiswal, A., & Kumar, P. (2025). *RAIL in the Wild: Operationalizing Responsible AI Evaluation Using Anthropic's Value Dataset*. arXiv. https://arxiv.org/html/2505.00204v1

15. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of NAACL-HLT (pp. 4171-4186).

16. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of machine learning research, 21(140), 1-67.

17. Pichai, S. (2023). *An important next step in our AI journey*. Google Blog. https://blog.google/intl/en-africa/products/explore-get-answers/an-important-next-step-on-our-ai-journey/

18. Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). *Lamda: Language models for dialog applications.* arXiv preprint arXiv:2201.08239.

19. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2023). *Palm: Scaling language modeling with pathways*. Journal of Machine Learning Research, 24(240), 1-113.

20. Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., ... & Wu, Y. (2023). *Palm 2 technical report*. arXiv preprint arXiv:2305.10403.

21. Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). *Gemini: a family of highly capable multimodal models*. *arXiv preprint arXiv:2312.11805*.

22. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). *Llama: Open and efficient foundation language models*. arXiv preprint arXiv:2302.13971.

23. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. arXiv preprint arXiv:2307.09288.

24. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... & Ganapathy, R. (2024). *The llama 3 herd of models*. arXiv e-prints, arXiv-2407.

25. Meta. (2024). Llama 3.1. *Llama.com*. https://www.llama.com/llama3_1/

26. Lee, J., Song, K. U., Yang, S., Lim, D., Kim, J., Shin, W., ... & Kim, T. H. (2025). *Efficient LLaMA-3.2-Vision by Trimming Cross-attended Visual Features*. arXiv preprint arXiv:2504.00557.

27. Meta. (2024). *Llama-3.3-70B-Instruct*. Hugging Face. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

28. Meta AI. (2025). *The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation*. Meta AI. https://ai.meta.com/blog/llama-4-multimodal-intelligence/

29. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W.E. (2023). *Mistral 7B.* ArXiv, abs/2310.06825.

30. Mistral AI. (2024). *Mistral Large*. Mistral AI. https://mistral.ai/news/mistral-large

31. Mistral AI. (2025). *Medium is the new large*. Mistral AI. https://mistral.ai/news/mistral-medium-3

32. Rastogi, A., Jiang, A. Q., Lo, A., Berrada, G., Lample, G., Rute, J., ... & Tang, Y. (2025). *Magistral*. arXiv preprint arXiv:2506.10910.

33. Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., ... & Zou, Y. (2024). *Deepseek llm: Scaling open-source language models with longtermism.* arXiv preprint arXiv:2401.02954.

34. Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., ... & Xu, Z. (2024). *Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.* arXiv preprint arXiv:2405.04434.

35. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., ... & Piao, Y. (2024). *Deepseek-v3 technical report.* arXiv preprint arXiv:2412.19437.

36. Sallam, M., Al-Mahzoum, K., Sallam, M., & Mijwil, M. M. (2025). *DeepSeek: Is it the end of generative AI monopoly or the mark of the impending doomsday?* Mesopotamian Journal of Big Data, 2025, 26-34.

37. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.* arXiv preprint arXiv:2501.12948.

38. Karpas, E., Abend, O., Belinkov, Y., Lenz, B., Lieber, O., Ratner, N., ... & Tenenholtz, M. (2022). *MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning*. arXiv preprint arXiv:2205.00445.

39. Lieber, O., Lenz, B., Bata, H., Cohen, G., Osin, J., Dalmedigos, I., ... & Shoham, Y. (2024). *Jamba: A hybrid transformer-mamba language model.* arXiv preprint arXiv:2403.19887.

40. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., ... & Yang, H. (2022). *Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.* In International conference on machine learning (pp. 23318-23340). PMLR.

41. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., ... & Qiu, Z. (2025). *Qwen3 technical report.* arXiv preprint arXiv:2505.09388.

42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need.* Advances in neural information processing systems, 30.

43. Derakhshani, M.M., Varghese, D., Fadaee, M., & Snoek, C.G. (2025). *NeoBabel: A Multilingual Open Tower for Visual Generation.* arXiv e-prints, arXiv-2507.

44. Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., ... & Hooker, S. (2025). *The leaderboard illusion*. arXiv preprint arXiv:2504.20879.

45. Dzikowski, R. (2025). *Grok 3: A Threat to Human-AI Interaction and Technological Control.* ResearchGate. https://www.researchgate.net/publication/389395726_Grok_3_A_Threat_to_Human-AI_Interaction_and_Technological_Control_licensed_under_CC_BY-SA_40

The article has been sent to the editors 23.09.25.
After processing 25.09.25.
Submitted for printing 30.09.25.

Copyright under license CCBY-SA4.0.