ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ

Навчально-науковий інститут економіки, управління, права та інформаційних технологій

Кафедра інформаційних систем та технологій

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття ступеня вищої освіти магістр

на тему: **«Аналіз архітектурних особливостей та функціональних можливостей передових моделей Reasoning-LLM»**

Виконав: здобувач вищої освіти
за освітньо-професійною програмою
Інформаційні управляючі системи
та технології
спеціальності 126 Інформаційні
системи та технології
освітнього ступеня магістр
групи 126ІСТ_мд_2024
Масич Андрій Леонідович
Керівник: Флегантов Леонід Олексійович
Рецензент: Ковальчук Станіслав Богданович

Полтава – 2025 року

ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ

Навчально-науковий інститут економіки, управління, права та інформаційних технологій

Кафедра інформаційних систем та технологій

Освітня програма Інформаційні управляючі системи та технології
Спеціальність 126 Інформаційні системи та технології
Рівень вищої освіти другий (магістерський)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Юрій УТКІН
«08» листопада 2024 року

З А В Д А Н Н Я
НА КВАЛІФІКАЦІЙНУ РОБОТУ ЗДОБУВАЧА ВИЩОЇ ОСВІТИ

Масича Андрія Леонідовича

1. Тема кваліфікаційної роботи: «Аналіз архітектурних особливостей та функціональних можливостей передових моделей Reasoning-LLM».

Керівник роботи к.ф.-м.н., доцент, професор кафедри інформаційних систем та технологій Флегантов Леонід Олексійович.

Затверджено наказом закладу вищої освіти від «31» жовтня 2025 року № 1332-ст

2. Строк подання здобувачем вищої освіти роботи «09» грудня 2025 р.

3. Вихідні дані до роботи: наукові публікації у рецензованих журналах, технічна документація OpenAI, Anthropic, DeepSeek, аналітичні звіти провідних лабораторій (Google DeepMind, Microsoft Research), матеріали з наукометричних баз Scopus, IEEE Xplore, arXiv, а також офіційних white-paper і тестових демонстрацій моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet.

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити):

Розділ 1. Теоретико-методологічні основи дослідження Reasoning-LLM

Розділ 2. Аналітичний огляд архітектурних особливостей та функціональних можливостей провідних моделей Reasoning-LLM

Розділ 3. Порівняльний аналіз та розроблення рекомендацій із застосування Reasoning-LLM

5. Перелік графічного матеріалу: схеми, рисунки за темою та об’єктом дослідження.

6. Консультанти розділів кваліфікаційної роботи:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання отримав
Оцінювання економічної ефективності результатів дослідження	Калініченко О. В., к. е. н., доцент, доцент кафедри економіки та публічного управління	24.11.2025	04.12.2025

7. Дата видачі завдання «08» листопада 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк виконання етапів кваліфікаційної роботи	Примітка
1	Вибір і затвердження теми роботи	29.10.2024 р.	
2	Складання і затвердження розгорнутого плану та завдання на кваліфікаційну роботу	30.10.2024 р. – 08.11.2024 р.	
3	Опрацювання джерел інформації	11.11.2024 р. – 27.12.2024 р.	
4	Збір, вивчення і обробка інформації, необхідної для виконання роботи	30.12.2024 р.– 19.01.2025 р.	
5	Виконання теоретико-методологічного розділу роботи	17.02.2025 р.– 16.05.2025 р.	
6	Виконання дослідницько-аналітичного розділу роботи	02.06.2025 р.– 13.07.2025 р.	
7	Виконання проектно-рекомендаційного розділу роботи	08.09.2025 р.– 14.11.2025 р.	
8	Розрахунок економічної ефективності результатів дослідження	24.11.2025 р.– 04.12.2025 р.	
9	Оформлення тексту роботи	05.12.2025 р.– 08.12.2025 р.	
10	Попередній захист роботи на кафедрі	09.12.2025 р.	
11	Доопрацювання роботи з урахуванням зауважень і пропозицій	10.12.2025 р. – 14.12.2025 р.	
12	Нормоконтроль	15.12.2025 р. – 16.12.2025 р.	
13	Захист кваліфікаційної роботи	18.12.2025 р.	

Здобувач вищої освіти **Андрій МАСИЧ**

Керівник роботи **Леонід ФЛЕГАНТОВ**

ПОЛТАВСЬКИЙ ДЕРЖАВНИЙ АГРАРНИЙ УНІВЕРСИТЕТ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ЕКОНОМІКИ, УПРАВЛІННЯ, ПРАВА ТА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ СИСТЕМ ТА ТЕХНОЛОГІЙ

МАСИЧ АНДРІЙ ЛЕОНІДОВИЧ

«АНАЛІЗ АРХІТЕКТУРНИХ ОСОБЛИВОСТЕЙ ТА ФУНКЦІОНАЛЬНИХ МОЖЛИВОСТЕЙ ПЕРЕДОВИХ МОДЕЛЕЙ REASONING-LLM»

Освітньо-професійна програма
Інформаційні управляючі системи та технології
Спеціальність 126 Інформаційні системи та технології
Ступінь вищої освіти Магістр

РЕФЕРАТ
кваліфікаційної роботи на здобуття кваліфікації –
магістр з інформаційних систем та технологій

Полтава – 2025 року

Кваліфікаційна робота складається зі вступу, 3 розділів, висновків, списку використаних джерел (58 найменувань), додатків. Робота містить 21 рисунок, 11 таблиць і викладена на 69 сторінках.

Основний зміст роботи

У роботі досліджено архітектурні та функціональні особливості сучасних reasoning-орієнтованих великих мовних моделей (Reasoning-LLM) на прикладі OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet. Актуальність теми зумовлена зростаючою потребою у використанні інтелектуальних систем, здатних не лише генерувати текст, а й здійснювати багатокрокове логічне міркування, аналіз складних контекстів і обґрунтування прийнятих рішень.

У першому розділі роботи розглянуто теоретичні засади розвитку reasoning-моделей, еволюцію трансформерних архітектур і принципи формування логічного мислення в LLM. Проаналізовано сучасні підходи до chain-of-thought, tree-of-thought, self-consistency та інші методи, що лежать в основі reasoning-LLM. Також визначено ключові архітектурні відмінності між класичними мовними моделями та моделями нового покоління.

Другий розділ присвячено детальному аналізу архітектурних і функціональних характеристик моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet. Досліджено особливості їхніх reasoning-механізмів, роботи з довгим контекстом, інтеграції зовнішніх інструментів, мультимодальності та підходів до самоперевірки результатів. Наведено порівняльні таблиці та схеми, що відображають відмінності у структурі, продуктивності та сфері застосування моделей.

У третьому розділі виконано порівняльний аналіз ефективності моделей у різних типах завдань: логічних, аналітичних, програмних, юридичних і бізнес-орієнтованих. Проведено експериментальні мікротести з використанням API, зокрема тестування роботи з довгим контекстом, аналіз видимості reasoning-кроків, оцінку затримки відповіді та стабільності результатів. На основі отриманих даних сформовано практичні рекомендації щодо вибору моделей залежно від умов застосування та ресурсних обмежень.

Окрему увагу приділено техніко-економічному обґрунтуванню, яке показало, що використання reasoning-моделей дозволяє скоротити трудомісткість аналітичних процесів, підвищити якість прийняття рішень і досягти позитивного показника ROI у корпоративних та дослідницьких системах. Запропоновано підхід до гібридного використання моделей, за якого сильні сторони однієї компенсують обмеження іншої.

Висновки

Метою кваліфікаційної роботи було дослідження архітектурних та функціональних особливостей сучасних reasoning-моделей і розробка рекомендацій щодо їх ефективного застосування в інформаційних системах. У ході дослідження встановлено, що OpenAI o1 забезпечує найвищу точність у завданнях складного логічного аналізу, DeepSeek R1 демонструє оптимальний баланс між продуктивністю та економічністю, а Claude 3.7 Sonnet є ефективним рішенням для мультимодальних і контекстно чутливих сценаріїв.

Порівняльний аналіз підтвердив, що reasoning-LLM є якісним етапом розвитку штучного інтелекту, оскільки поєднують мовне моделювання з логічним обґрунтуванням результатів. Результати роботи можуть бути використані при розробці інтелектуальних агентів, аналітичних систем, освітніх платформ і корпоративних інформаційних рішень.

Наукова новизна роботи полягає у систематизованому порівнянні сучасних reasoning-моделей з урахуванням архітектурних, функціональних і економічних показників, а практична цінність – у розробці рекомендацій та демонстраційних прикладів застосування моделей у реальних сценаріях.

Список публікацій здобувача

1. Масич А. Досвід побудови захищеної інфраструктури GPON та аналіз технології Reasoning-LLM. *Матеріали науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р.* Вип. XI. Полтава: ПДАУ, 79 с. С. 69-72.
2. Масич А. Застосування Reasoning-LLM у різних галузях. *Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р.* Полтава: ПДАУ, 2025 р. 120 с. С. 68-73.
3. Флегантов Л., Масич А. Основні принципи та архітектури Reasoning-LLM. *Advanced Technologies in Scientific Research: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical* Conference. International Scientific Unity. November 19-21, 2025. Rotterdam, Netherlands. 210-215 p. URL: https://isu-conference.com/en/archive/advanced-technologies-in-scientific-research-19-11-25/ (дата звернення: 20.11.2025).
4. Флегантов Л., Масич А., Левченко Ю. Архітектурні та функціональні особливості провідних моделей Reasoning-LLM. *Progressive Approaches in*

Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference. International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark. 338-344 p. URL: https://isu-conference.com/arkhiv/progressive-approaches-in-science-and-engineering-26-11-25/ (дата звернення: 27.11.2025).

АНОТАЦІЯ

Масич А.Л. «Порівняльний аналіз архітектурних та функціональних особливостей сучасних reasoning-моделей». Кваліфікаційна робота на правах рукопису. Кваліфікаційна робота на здобуття ступеня вищої освіти магістр за освітньо-професійною програмою Інформаційні системи та технології, спеціальність 126. Полтавський державний аграрний університет, Полтава, 2025.

Робота присвячена дослідженню сучасних reasoning-орієнтованих великих мовних моделей та аналізу їхньої ефективності у складних аналітичних і когнітивних завданнях. Проведено порівняльний аналіз моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet, визначено їхні сильні та слабкі сторони, а також сформовано рекомендації щодо практичного застосування.

Ключові слова: reasoning-LLM, OpenAI o1, DeepSeek R1, Claude 3.7 Sonnet, логічне міркування, великі мовні моделі, штучний інтелект.

ANNOTATION

Masich A.L. «Comparative Analysis of Architectural and Functional Features of Modern Reasoning Models". Master's thesis manuscript.

The thesis is devoted to the study of modern reasoning-oriented large language models and the analysis of their effectiveness in complex analytical and cognitive tasks. A comparative analysis of OpenAI o1, DeepSeek R1, and Claude 3.7 Sonnet is conducted, highlighting their architectural differences, functional capabilities, and application domains. Practical recommendations for model selection and usage are proposed based on experimental evaluation.

Keywords: reasoning LLM, OpenAI o1, DeepSeek R1, Claude 3.7 Sonnet, logical reasoning, large language models, artificial intelligence.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАК

ШІ – штучний інтелект

API (Application Programming Interface) – програмний інтерфейс прикладного програмування

AWS (Amazon Web Services) – хмарна платформа Amazon Web Services

CoT (Chain-of-Thought) – ланцюжок міркувань

DMU (Dynamic Memory Unit) – модуль динамічної пам’яті

GPU (Graphics Processing Unit) – графічний процесор

IoT (Internet of Things) – інтернет речей

LLM (Large Language Model) – велика мовна модель

MATH – тестовий набір задач з математики для оцінювання моделей

RAG (Retrieval-Augmented Generation) – генерація з підкріпленням пошуком

RLHF (Reinforcement Learning from Human Feedback) – навчання з підкріпленням на основі людського зворотного зв’язку

STEM (Science, Technology, Engineering and Mathematics) – наука, технології, інженерія та математика

ToT (Tree-of-Thought) – деревоподібна структура міркувань

Reasoning-LLM (Reasoning Large Language Model) – велика мовна модель із підтримкою логічного міркування

ВСТУП

Стрімкий розвиток штучного інтелекту упродовж останніх років призвів до появи нового покоління великих мовних моделей, здатних не лише генерувати текст, а й виконувати складні логічні міркування, аналізувати контекст, будувати багатокрокові ланцюжки висновків і демонструвати елементи узагальненого мислення. Такі моделі отримали назву Reasoning-LLM – мовні моделі з розвиненими механізмами логічного та причинно-наслідкового мислення. Вони становлять новий еволюційний етап розвитку LLM після моделей типу GPT-4, Claude 3 та LLaMA 3, пропонуючи не лише потужніші параметричні архітектури, а й принципово нові механізми «CoT», «ToT», «self-verification» і «reflection-based reasoning». Завдяки цим підходам Reasoning-LLM показують істотно вищу якість у завданнях математичного доведення, програмного аналізу, аналітики даних, логічного аргументування та автоматизованого наукового пошуку.

На сьогодні провідними представниками цього напрямку є OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet – моделі, що втілюють різні підходи до організації reasoning-процесів. Модель OpenAI o1 орієнтована на «мислення перед відповіддю» (think-before-respond), DeepSeek R1 інтегрує багаторівневу систему внутрішнього міркування з елементами рефлексії, а Claude 3.7 Sonnet поєднує причинно-наслідкову логіку з контекстно-чутливим розумінням природної мови. Вивчення архітектурних особливостей і порівняння функціональних можливостей цих моделей має надзвичайну наукову та практичну значущість, адже воно дозволяє зрозуміти принципи побудови систем нового покоління, які у найближчі роки визначатимуть розвиток ШІ-галузі.

Актуальність теми даної роботи зумовлена потребою у комплексному аналізі архітектур, методів навчання та reasoning-механізмів сучасних LLM, а також у визначенні їх ефективних напрямів застосування у завданнях, що потребують високого рівня логічного висновування. Дослідження спрямоване на виявлення переваг, обмежень і потенціалу розвитку Reasoning-LLM, що є

важливим для науковців, розробників та аналітиків, які створюють інтелектуальні системи наступного покоління.

Зв'язок роботи з науковими програмами, планами, темами. Магістерська кваліфікаційна робота відповідає дослідженням в межах науково-дослідної ініціативної тематики «Організаційно-методологічні аспекти впровадження інформаційно-комунікаційних систем і технологій в управлінні діяльністю сучасних організацій та підприємств за умов переходу до цифрової економіки» (ДРН 0123U105060, 2023-2028 рр.), що реалізується на кафедрі інформаційних систем та технологій, тематиці навчально-дослідної лабораторії досліджень інтелектуальних систем, комп'ютерних мереж інтернет речей кафедри інформаційних систем та технологій Полтавського державного аграрного університету.

Мета роботи полягає у здійсненні комплексного аналізу архітектурних особливостей і функціональних можливостей передових Reasoning-LLM – OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet – із визначенням їх переваг, недоліків і сфер найефективнішого застосування.

Для досягнення поставленої мети були поставлені такі *завдання*:

- дослідити теоретико-методологічні основи побудови та функціонування Reasoning-LLM;
- проаналізувати архітектурні принципи та внутрішні механізми reasoning-моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet;
- здійснити порівняльний аналіз функціональних можливостей зазначених моделей за критеріями продуктивності, точності, контекстного розуміння та ефективності reasoning-процесів;
- виявити сильні та слабкі сторони кожної моделі;
- розробити практичні рекомендації щодо вибору та застосування Reasoning-LLM у різних галузях;
- провести техніко-економічне обґрунтування отриманих результатів.

Об'єкт дослідження – архітектури та принципи функціонування сучасних Reasoning-LLM.

Предмет дослідження – архітектурні особливості та функціональні можливості моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet.

Методи дослідження. У процесі виконання магістерської роботи застосовано такі методи: аналітичний огляд наукових і технічних джерел для визначення сучасних тенденцій розвитку Reasoning-LLM; системний аналіз архітектур, механізмів reasoning, методів навчання та оптимізації моделей; порівняльний аналіз продуктивності та функціональних можливостей моделей за головними критеріями; експериментальні дослідження на тестових завданнях, що потребують багатокрокового логічного міркування; узагальнення та інтерпретація отриманих результатів для формування практичних рекомендацій.

Інформаційна база дослідження складається з наукових публікацій у рецензованих журналах, технічної документації OpenAI, Anthropic, DeepSeek, аналітичних звітів провідних лабораторій (Google DeepMind, Microsoft Research), матеріалів із наукометричних баз Scopus, IEEE Xplore, arXiv, а також офіційних white-paper і тестових демонстрацій моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet.

Елементи наукової новизни полягають у: систематизації сучасних підходів до побудови reasoning-архітектур великих мовних моделей; проведенні порівняльного аналізу reasoning-механізмів моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet; виявленні головних архітектурних рішень, що впливають на ефективність логічного висновування та інтерпретації контексту; розробленні рекомендацій щодо вибору оптимальних моделей для завдань, пов’язаних із багатокроковим міркуванням; формуванні узагальненої аналітичної моделі, що відображає співвідношення між архітектурною складністю, reasoning-потужністю та ефективністю використання обчислювальних ресурсів.

Практична значущість отриманих результатів полягає у можливості їх застосування для: розроблення аналітичних, освітніх та дослідницьких систем, які потребують автоматизованого логічного висновування; побудови програмних агентів і систем підтримки прийняття рішень на основі Reasoning-LLM; підвищення ефективності використання обчислювальних ресурсів при

впровадженні reasoning-моделей у корпоративні середовища; подальшого розвитку українських ініціатив у сфері створення власних моделей штучного інтелекту reasoning-типу.

Апробація результатів дослідження. За результатами роботи опубліковано тези доповідей: «Досвід побудови захищеної інфраструктури GPON та аналіз технології Reasoning-LLM». Матеріали науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р. Вип. XI. Полтава: ПДАУ, 2025; «Застосування Reasoning-LLM у різних галузях». Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р. Полтава: ПДАУ, 2025.; «Основні принципи та архітектури Reasoning-LLM». Advanced Technologies in Scientific Research: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference. International Scientific Unity. November 19-21, 2025. Rotterdam, Netherlands, 2025.; «Архітектурні та функціональні особливості провідних моделей Reasoning-LLM». Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical Conference. International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark, 2025.

Структура та обсяг кваліфікаційної роботи. Магістерська робота складається зі вступу, трьох розділів, висновків, списку використаних джерел і додатків. Основний текст викладено на 75 сторінках, містить 21 рисунок і 11 таблиць. Список використаних джерел налічує 58 найменувань.

РОЗДІЛ 1
ТЕОРЕТИКО-МЕТОДОЛОГІЧНІ ОСНОВИ ДОСЛІДЖЕННЯ REASONING-LLM

1.1 Основні принципи та архітектури Reasoning-LLM

У сучасній парадигмі розвитку штучного інтелекту спостерігається перехід від класичних великих мовних моделей (LLM), орієнтованих переважно на статистичні закономірності текстів, до моделей нового покоління – Reasoning-LLM, що орієнтовані не лише на генерацію зв'язного тексту, а й на виконання складного логічного міркування, контекстного аналізу та дедуктивного висновку.

На рівні базових характеристик різницю між цими двома класами моделей можна узагальнити за кількома критеріями: мета використання, тип навчальних даних, характер процесу відповіді, прозорість міркування, ресурсоємність та сфера найефективнішого застосування. Класичні LLM зосереджені на прогнозуванні наступного токена і показують високу швидкість та низьку вартість за один запит, тоді як Reasoning-LLM оптимізуються для задач «другого рівня» – доказових, аналітичних і планувальних, де важливі аргументованість і логічна послідовність результату.

Поява Reasoning-LLM стала результатом еволюції трансформерних архітектур, що з моменту свого впровадження заклали основи сучасного напрямку глибинного навчання у природній мові [1]. Саме трансформер став базовою архітектурою, на якій надалі базуються reasoning-модулі, зовнішні інструменти та механізми пояснюваності.

Навчання Reasoning-LLM потребує не просто великого корпусу текстів, як для звичайних LLM, а таких даних, що містять складні задачі та їх покрокові розв'язання. Це принципово змінює вимоги до навчальних наборів: від «масового» тексту – до структурованих прикладів міркувань, які відображають ланцюги думок, проміжні розрахунки, варіанти рішень і їх порівняння. Така

організація навчання дозволяє моделі «думати довше», генеруючи внутрішні токени міркування і розбиваючи задачу на етапи. На відміну від попередніх LLM, у яких генерація мовлення була безпосереднім наслідком ймовірнісного розподілу слів, Reasoning-LLM володіють властивістю створювати внутрішні ланцюги роздумів, фіксувати їх у прихованих станах або видимому CoT та на цій основі формулювати кінцеву відповідь. Основні відмінності між звичайними LLM та Reasoning-LLM узагальнені у таблиці 1.1.

Таблиця 1.1 – Основні відмінності між LLM та Reasoning-LLM

Критерій	Класична LLM	Reasoning-LLM
Основна мета	Генерація зв'язного, граматично правильного та контекстуально релевантного тексту. Прогнозування наступного токена (слова).	Логічне розв'язання проблєм, планування, дедуктивний висновок.
Навчання	Великий обсяг загального тексту.	Важливий не лише обсяг, а наявність складних задач та їх покрокових рішень у навчальних даних.
Процес відповіді	Відбувається за одну фазу генерації; модель швидко видає результат.	Часто може розбивати завдання на етапи, робити проміжні розрахунки, порівнювати альтернативи, «думати довше», генеруючи більше внутрішніх токенів (токенів міркування).
Прозорість	Етап міркування відсутній або невидимий. Ми бачимо лише кінцевий результат.	Деякі моделі можуть на вимогу показувати свої роздуми/кроки (ланцюг думок, CoT), що спрощує налагодження та розуміння відповіді.
Ресурси та час	Швидкі та відносно дешеві.	Повільніші та дорожчі (потребують більше обчислювальних зусиль та часу для складних завдань).
Сильні сторони	Швидкість, низька вартість (на запит), відтворення знань, узагальнення інформації, задачі «першого рівня» (наприклад, чат-боти, написання тексту, підсумовування).	Точність, обґрунтованість, задачі «другого рівня» (логічні докази, аналіз, планування, багатоетапне прийняття рішень).

Архітектура Reasoning-LLM передбачає поєднання двох рівнів когнітивної обробки запитів:

– нижній («мовний») рівень відповідає за розпізнавання та генерацію тексту, коду або інших послідовностей;

– верхній (логічний) рівень реалізує процеси мислення, планування та прийняття рішень, оперує проміжними кроками та внутрішніми представленнями задачі.

На рисунку 1.1 схематично показано базову структуру Reasoning-LLM: вхідний запит користувача (User) обробляється мовним ядром моделі, після чого активується модуль міркування, здатний виконувати аналіз задачі, генерувати проміжні кроки та, за потреби, звертатися до зовнішніх джерел знань.

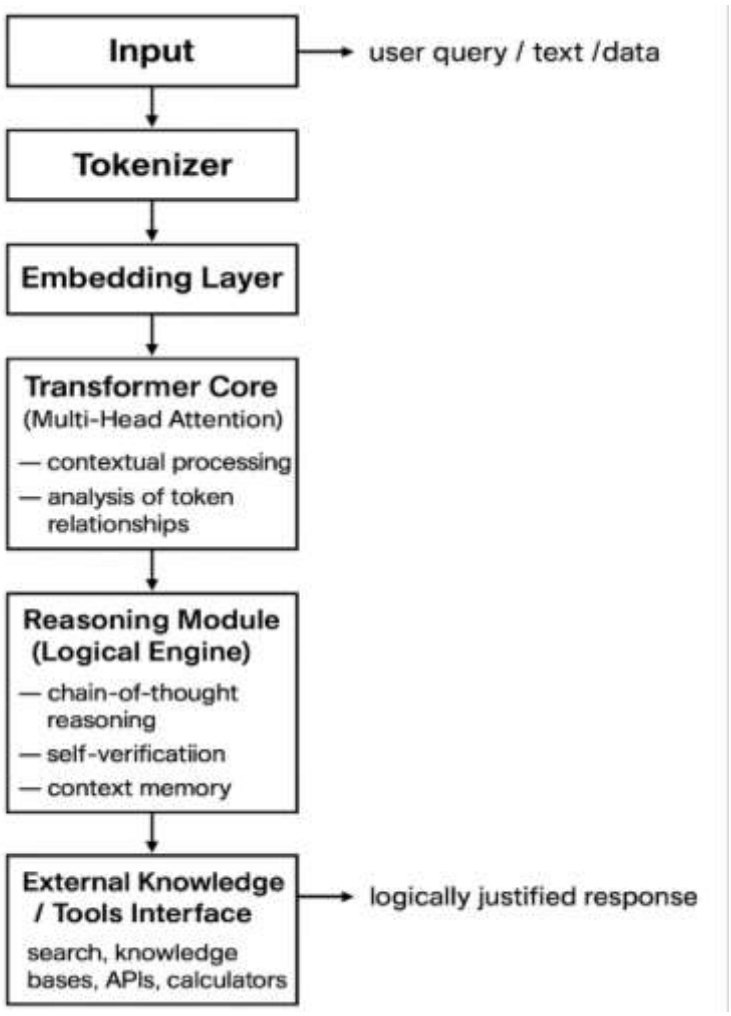

Рисунок 1.1 – Загальна схема архітектури Reasoning-LLM

Такий поділ на мовний та reasoning-рівні відображає перехід від суто статистичного моделювання мови до моделювання процесів мислення.

Основною відмінністю Reasoning-LLM від попередніх генерацій LLM (таких як GPT-3 або PaLM) є орієнтація не лише на предиктивне продовження

тексту, а на логічне розв’язання проблєм. Якщо традиційні LLM моделювали лише поверхневу структуру мови, то reasoning-моделі намагаються імітувати когнітивні процедури: постановку підзадач, побудову проміжних аргументів, використання зовнішніх інструментів (калькуляторів, пошукових систем, баз знань) та планування багатокрокових дій. Такий підхід безпосередньо спрямований на розв’язання проблєми «галюцинацій» і відсутності прозорості, оскільки модель може не лише надати відповідь, а й показати, як вона до неї дійшла.

З методологічного погляду Reasoning-LLM поєднують нейронні мережеві підходи із когнітивними принципами штучного мислення. В основі архітектури залишається трансформер із механізмом самоуваги (self-attention), який аналізує складні залежності в контексті. У новітніх реалізаціях таких моделей, впроваджено низку покращених компонентів, серед яких:

– розширене контекстне вікно, що дозволяє моделі утримувати в пам’яті значні обсяги інформації (десятки або сотні тисяч токенів);

– механізми зовнішньої пам’яті RAG, які забезпечують доступ до баз знань і документів;

– покрокове міркування (step-by-step reasoning), що дозволяє моделі обґрунтовувати власні висновки;

– інструментальне розмірковування (tool-augmented reasoning), коли модель здатна використовувати зовнішні сервіси для розрахунків, символьних перетворень або пошуку інформації.

У цьому контексті центральним елементом архітектури Reasoning-LLM виступає модуль міркування (Reasoning Module), який можна інтерпретувати як агент або планувальник. Він аналізує отримане завдання, вирішує, або достатньо внутрішніх знань моделі, або потрібне звернення до зовнішніх інструментів, порівнює альтернативні шляхи розв’язання та формує узгоджену стратегію відповіді. Ядро LLM при цьому використовується переважно як мовний інтерфейс, що формулює підсумковий текст, тоді як логічна обробка відбувається у reasoning-блоці (рисунок 1.2).

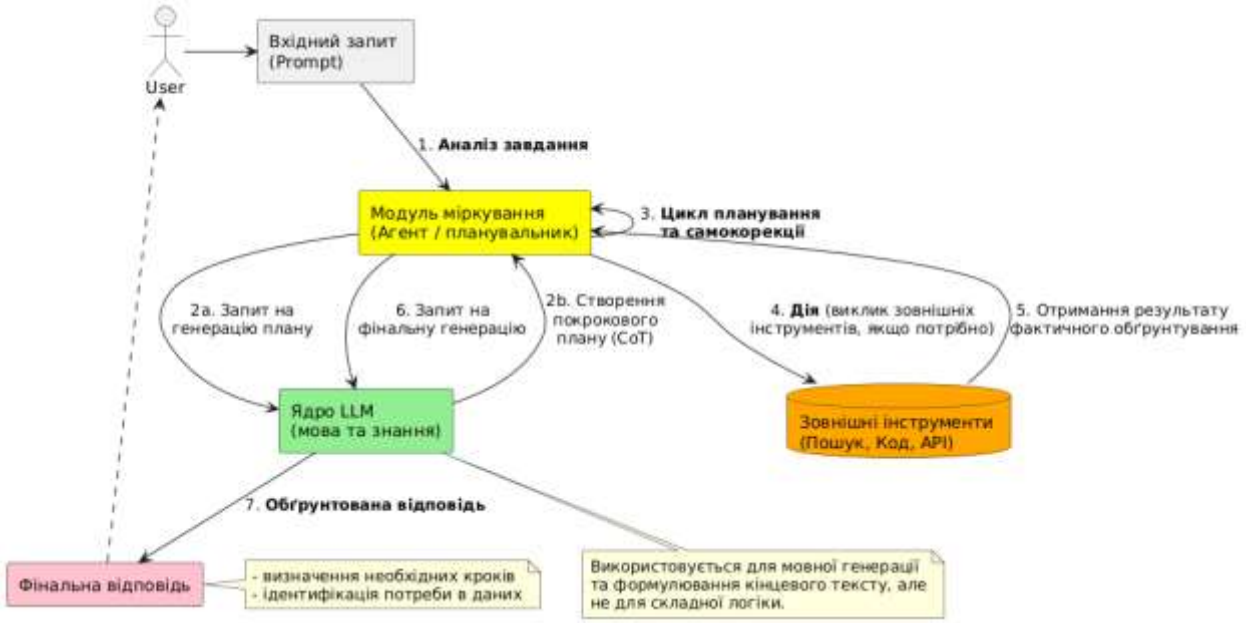

Рисунок 1.2 – Спрощена схема архітектури Reasoning-LLM

Наведена вище спрощена схема показує, що Reasoning-LLM функціонує як гібридна система:

– на вхід надходить запит користувача;

– reasoning-модуль аналізує його, за потреби звертається до зовнішніх інструментів (бази знань, пошук, калькулятори, символьні розв’язувачі);

– після цього сформований план рішення передається до мовного ядра, яке генерує фінальний текст відповіді.

Такий підхід дає змогу чітко розділити задачі «міркування» і «формулювання», що підвищує прозорість та керованість системи.

Аналіз сучасних наукових публікацій (у тому числі оглядів, розміщених у arXiv preprints за 2023–2025 рр.) свідчить, що архітектура reasoning-моделей поступово набуває рис гібридних систем, де поєднуються нейромережеві та символьні компоненти.

Такий напрям розвитку спрямований на подолання головного обмеження класичних LLM – «чорної скриньки», тобто відсутності прозорих механізмів логічного виведення.

Розробники прагнуть створити системи, які не лише дають правильні результати, а й здатні пояснити хід власних міркувань (рисунок 1.3).

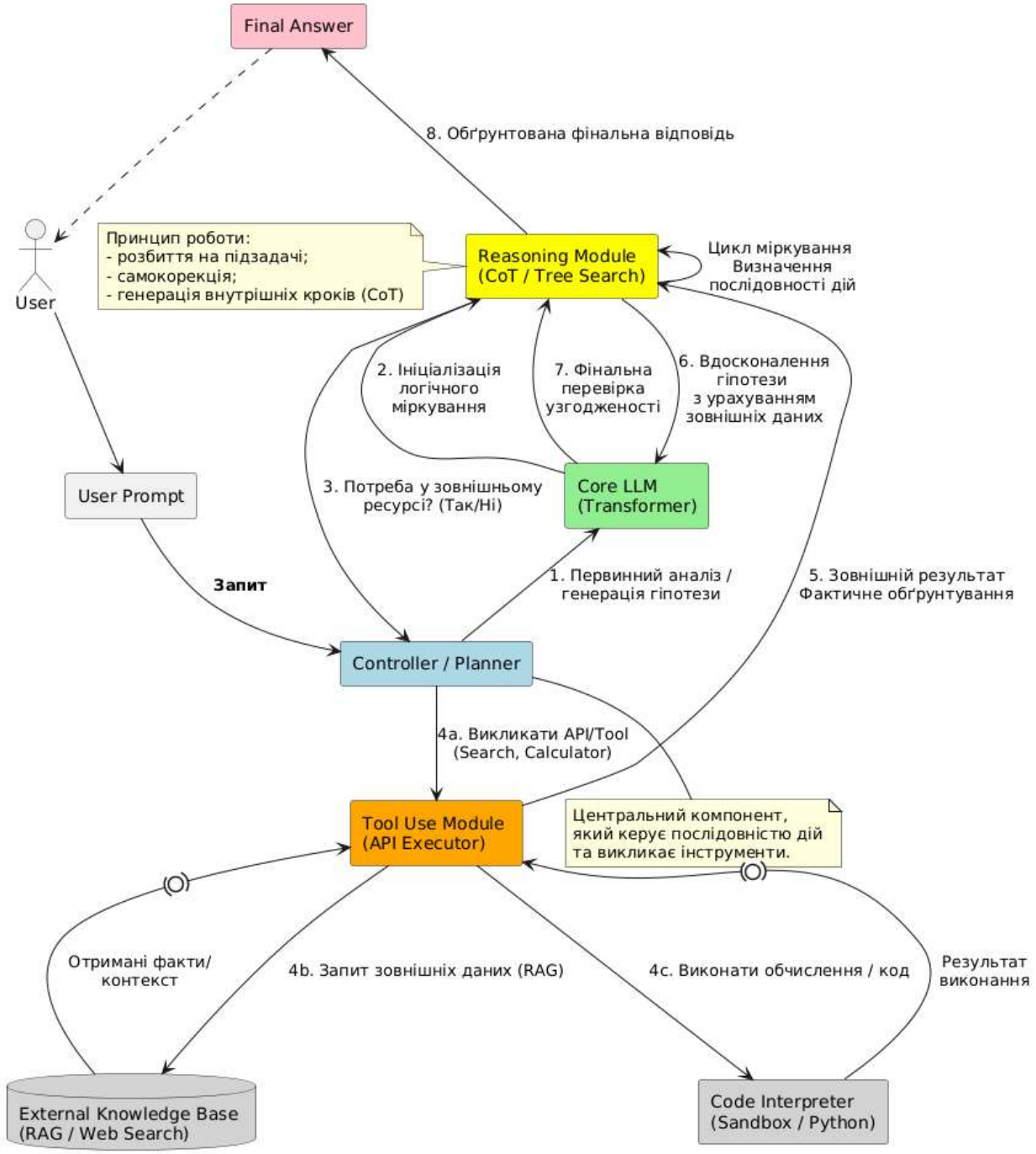

Рисунок 1.3 – Деталізована схема архітектури Reasoning-LLM

Деталізована схема рисунока 1.3 розкриває внутрішній механізм роботи Reasoning-LLM: від попередньої обробки вхідних даних, токенізації та побудови ембеддингів – до проходження через трансформерне ядро, активації reasoning-блоку, взаємодії з зовнішніми інструментами й формування вихідної відповіді. У цьому представленні добре видно, що Reasoning-LLM – це не просто «велика мережа», а багаторівнева система з чітко розподіленими функціями [4, 5].

Згідно з оглядом, опублікованим у arXiv preprints, формування reasoning-здібностей у LLM пов’язане не лише з архітектурними покращеннями, а й із методами навчання, що стимулюють модель «мислити» [2]. Застосовуються такі техніки, як CoT [2], Self-Consistency Decoding [6], ToT [3] та RLHF [4]. Вказані підходи є головними механізми, що перетворюють трансформер з «мовної» моделі до системи, здатної формувати послідовні, багатокрокові ланцюги міркувань. На основі даних сучасних досліджень можна зробити висновок, що архітектури Reasoning-LLM є результатом поступового синтезу трансформерних технологій (self-attention, масштабовані контекстні вікна), когнітивних підходів (моделювання ланцюгів міркування, планування, абстрагування) та специфічних методів навчання, орієнтованих на формування логічних висновків засобами Reasoning-LLM (CoT, ToT, RLHF тощо).

Таким чином, поява Reasoning-LLM – це новий етап розвитку штучного інтелекту, у якому головну роль відіграє не лише обсяг параметрів моделі або обчислювальні потужності, а здатність системи до раціонального аналізу, аргументації та пояснення своїх рішень. Перспективи подальших досліджень у цій сфері пов’язані з удосконаленням архітектур мовних моделей, інтеграцією символьних і нейронних підходів, а також розробкою методів контролю достовірності reasoning-висновків [7].

1.2 Методи навчання та оцінки Reasoning-LLM

Ефективність та інтелектуальна спроможність нейромережевих моделей класу Reasoning-LLM значною мірою визначаються методами їх навчання та оцінювання. На відміну від класичних LLM, що головним чином орієнтовані на відтворення мовних закономірностей, reasoning-моделі навчають мислити – тобто оперувати проміжними кроками, аргументами та внутрішніми міркуваннями, що ведуть до логічно обґрунтованої відповіді.

Методологічні засади навчання таких систем ґрунтуються на поєднанні кількох підходів: масштабного передтренування на текстових корпусах, цілеспрямованого донавчання на reasoning-завданнях, а також підкріплення моделі зворотним зв’язком від людини або зовнішніх критеріїв якості. У цьому контексті можна виокремити три основні рівні навчання Reasoning-LLM: предтренування (pretraining), інструкційне навчання (instruction fine-tuning) та підкріплювальне навчання (reinforcement learning).

Початковий етап передбачає навчання моделі на гігантських текстових масивах, що охоплюють різноманітні джерела – наукові статті, енциклопедії, коди, публіцистику, форуми тощо. Цей процес забезпечує формування мовного базису, що дозволяє системі оперувати граматичними структурами, синтаксичними залежностями та статистичними патернами.

Однак сам по собі цей етап не створює здатності до міркування. Як відзначено у [8], традиційне передтренування створює «мовний інстинкт», але не «логічний розум». Тому наступні етапи навчання спрямовані на формування когнітивної компетенції моделі – уміння послідовно мислити, аргументувати та приймати рішення.

Етап інструкційного навчання базується на спеціально підготовлених наборах даних, де моделі демонструють, як саме слід розмірковувати. Формат таких наборів передбачає інструкції та приклади розв’язання задач із проміжними кроками. Одним із найвідоміших методів інструкційного навчання є CoT Fine-Tuning, уперше запропонований у роботі [3].

Сутність методу полягає у тому, що модель навчається не лише давати кінцеву відповідь, а й пояснювати процес її отримання, створюючи «ланцюжок думок». Це перетворює LLM на reasoning-машину, що відтворює логіку людського мислення.

У подальшому цю ідею розвинули методи ToT [4] та Graph-of-Thoughts [9], які дозволяють моделі формувати дерево альтернативних міркувань і вибирати найоптимальнішу гілку (рисунок 1.4).

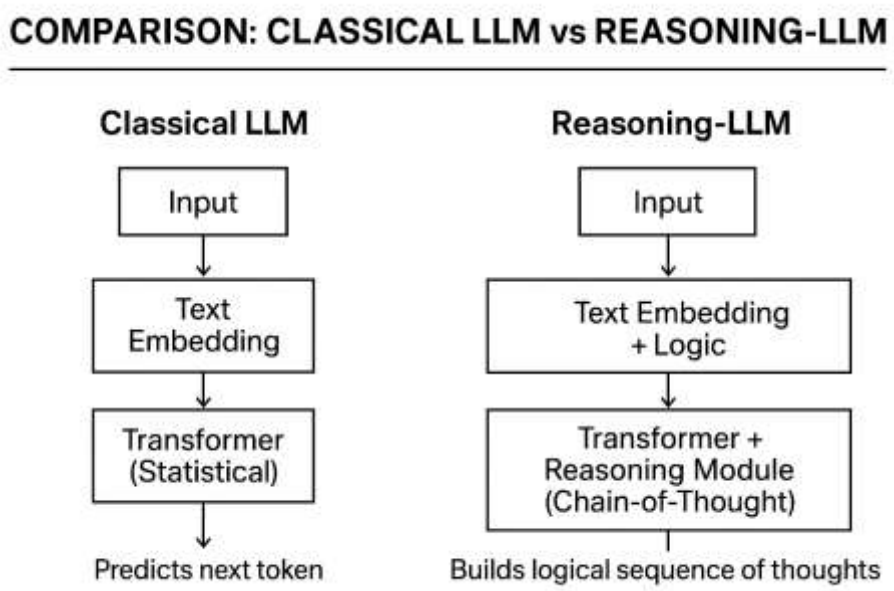

Рисунок 1.4 – Порівняння класичної LLM і Reasoning-LLM

До інструкційного навчання належать також методи Self-Consistency Decoding, що передбачають генерацію кількох ланцюгів роздумів для однієї задачі з подальшим вибором найузгодженішої відповіді [6]. Такий підхід не лише підвищує точність рішень, але й формує більш стабільну логічну поведінку моделі.

Одним із основних методів, що дозволяє узгодити поведінку моделі з людськими цінностями й логікою, є RLHF. Уперше цей підхід був масштабно реалізований у GPT-3.5 і GPT-4 від OpenAI [10]. У процесі RLHF модель генерує кілька варіантів відповідей, після чого експерти-оцінювачі визначають, який із них є найкращим. Ці оцінки використовуються для навчання reward-моделі, яка потім керує подальшим оновленням основної LLM за принципом підкріплювального навчання. У reasoning-моделях RLHF відіграє особливо важливу роль, оскільки дозволяє відфільтровувати поверхневі або «галюцинаторні» відповіді, надаючи перевагу логічно послідовним.

На наступному етапі з’явилися модифікації цього підходу – RLAIF (Reinforcement Learning from AI Feedback), коли роль людських експертів виконує вже інша, більш стабільна модель (наприклад, GPT-4). Це дає змогу масштабувати навчання reasoning-систем без значних людських витрат.

Останні розробки у сфері Reasoning-LLM передбачають створення самонавчальних систем, які здатні аналізувати власні помилки, формувати нові навчальні приклади й покращувати свої міркування. Такий підхід реалізовано у моделях Self-Reflective LLM, де модель генерує пояснення до власних відповідей, оцінює їх правильність і оновлює внутрішні параметри без участі людини [11].

Подібні механізми поєднують елементи метакогніції – здатності «мислити про власне мислення», що вважається однією з головних рис розуму. Це відкриває шлях до створення автономних reasoning-агентів, які не просто виконують завдання, а й удосконалюють свої стратегії розмірковування.

Оцінювання reasoning-моделей є складним завданням, оскільки воно має враховувати не лише точність кінцевих відповідей, а й якість процесу міркування. Традиційні метрики (наприклад, BLEU, ROUGE, Perplexity) виявилися недостатніми для аналізу reasoning-здатності, тому було розроблено низку спеціалізованих тестів [12]:

- benchmark-и логічного мислення – такі, як BIG-Bench Hard, GSM8K, ARC-Challenge, MATH та DROP – оцінюють здатність моделі до арифметичних, логічних і абстрактних міркувань;
- Evaluating CoT Coherence (ECoTC) – нові методи, що аналізують внутрішні кроки міркування моделі, перевіряючи їх логічну послідовність;
- Human Preference Alignment Metrics (HPA) – метрики, що визначають ступінь узгодженості відповідей моделі з людським судженням;
- Explainability and Faithfulness Metrics (EFM) – оцінюють, наскільки пояснення, надані моделлю, дійсно відображають процес її прийняття рішення, а не є вигаданими постфактум.

Існують також комбіновані методики оцінювання якості моделей, що об’єднують автоматизовані обчислювальні метрики з експертною людською оцінкою. Такий підхід дозволяє не лише вимірювати якість reasoning, а й відстежувати тенденції деградації логічного ланцюга (phenomenon of reasoning

drift), що виникає при збільшенні розміру моделі або змінах у навчальному корпусі.

Таким чином, методи навчання та оцінювання Reasoning-LLM формують багаторівневу систему, у якій поєднуються статистичне передтренування, когнітивно орієнтоване донавчання та підкріплення через людський або штучний зворотний зв’язок. Їх ефективність залежить не лише від обсягу даних, а й від структури навчальних завдань, які мають відображати реальні процеси мислення.

Сучасна тенденція розвитку у цій сфері полягає в переході від моделей, що просто відтворюють мову, до систем, здатних самостійно формувати, перевіряти та вдосконалювати власні міркування. Це створює фундамент для наступного покоління штучного інтелекту – моделей, які не просто імітують розум, а наближаються до його когнітивної сутності.

1.3 Застосування Reasoning-LLM у різних галузях

Як зазначалося вище, Reasoning-LLM – це моделі, здатні не лише генерувати текст, а й виконувати логічні міркування, що є основою їх практичного застосування. Завдяки цьому вони набувають особливого значення у сферах, де потрібне логічне узагальнення та інтерпретація неповних даних, – від медичних експертних систем і аналітичних центрів до наукових лабораторій та автоматизованих систем керування.

Сьогодні основними сферами застосування Reasoning-LLM є медицина, право, освіта, бізнес-аналітика, проєктний менеджмент, технічне проєктування, державне управління, наукові дослідження, а також сфера кібербезпеки. Кожна з цих галузей має специфічні потреби, для задоволення яких традиційні LLM часто виявляються недостатніми [13,].

Застосування Reasoning-LLM у різних галузях представлене на рисунку 1.5.

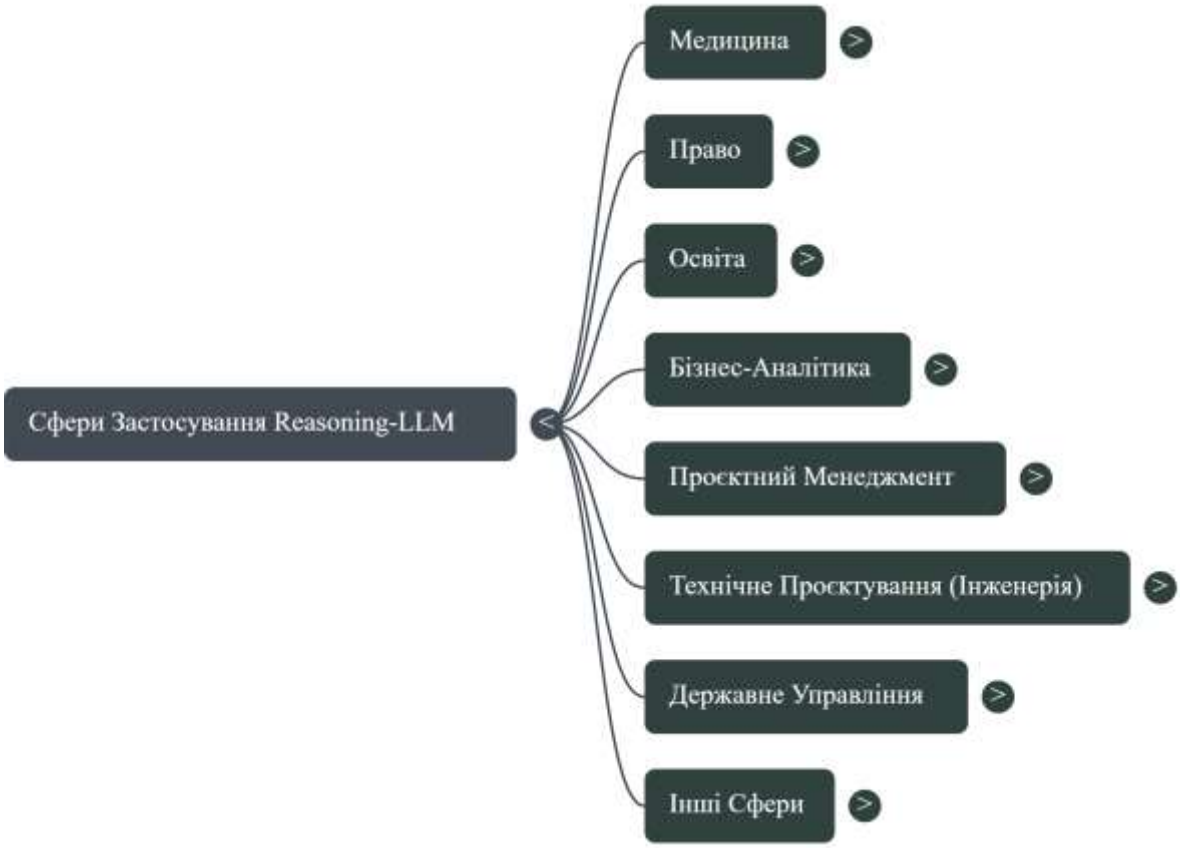

Рисунок 1.5 – Застосування Reasoning-LLM у різних галузях

У медицині Reasoning-LLM застосовуються для аналізу електронних медичних записів, інтерпретації результатів обстежень і формування попередніх діагнозів на основі анамнезу (рисунок 1.6)

Рисунок 1.6 – Приклади застосування Reasoning-LLM у медицині

Завдяки здатності до логічного міркування моделі можуть зіставляти клінічні симптоми з діагностичними шаблонами, формувати гіпотези й

обґрунтовувати варіанти лікування. Це сприяє підвищенню якості клінічного мислення та зменшенню ймовірності діагностичних помилок у складних випадках. На практиці Reasoning-LLM уже використовуються для аналізу онкологічних даних, підбору таргетної терапії та формування індивідуальних терапевтичних планів [14].

У юридичній практиці Reasoning-LLM здатні ефективно обробляти великі масиви нормативної документації, структурувати їх за релевантністю, формувати юридичні висновки, виявляти суперечності в контрактах і позовах, а також автоматизувати складання правових документів. Reasoning-LLM можуть пояснювати логіку ухвалених рішень, співвідносити їх із чинним законодавством і судовою практикою, що робить їх корисним інструментом для попереднього аналізу справ і підготовки процесуальних документів. Приклади застосування Reasoning-LLM у правознавстві представлені на рисунку 1.7.

Рисунок 1.7 – Приклади застосування Reasoning-LLM у правознавстві

В освітній сфері Reasoning-LLM використовуються для створення адаптивного навчального контенту, генерації логічних задач, формування варіативних тестових завдань і автоматизованого оцінювання студентських відповідей з урахуванням аргументації та ходу мислення та для іншого, що стосується навчалдьного процесу. Модель здатна не лише перевірити правильність відповіді, а й відтворити типову логіку міркування, виявити помилки у розумуванні та надати індивідуальний зворотний зв’язок. Це сприяє

розвитку критичного мислення й формуванню в здобувачів освіти навичок обґрунтованого судження [15]. Приклади застосування Reasoning-LLM в освіті представлені на рисунку 1.8.

Рисунок 1.8 – Приклади застосування Reasoning-LLM в освіті

У бізнес-аналітиці Reasoning-LLM застосовуються для побудови складних прогнозних моделей, аналізу зв’язків між головними бізнес-показниками та прийняття стратегічних рішень на основі виявлених закономірностей. Вони здатні виявляти приховані ризики, моделювати сценарії розвитку ринку, прогнозувати наслідки управлінських рішень, а також формувати обґрунтовані рекомендації щодо оптимізації бізнес-процесів. Приклади застосування Reasoning-LLM у бізнес-аналітиці представлені на рисунку 1.9.

Рисунок 1.9 – Приклади застосування Reasoning-LLM у бізнес-аналітиці

У проєктному менеджменті Reasoning-LLM забезпечують підтримку повного життєвого циклу проєкту: від формування цілей і декомпозиції завдань до контролю виконання та оцінювання ризиків. Вони можуть аналізувати залежності між задачами, пропонувати раціональний розподіл ресурсів, оптимізувати календарні плани та виявляти потенційні конфлікти на ранніх

етапах планування. У такому форматі Reasoning-LLM виступають як «віртуальні асистенти керівника проєкту», які моделюють сценарії розвитку подій і пропонують обґрунтовані альтернативи [16]. Застосування Reasoning-LLM у проєктному менеджменті представлене на рисунку 1.10.

Рисунок 1.10 – Застосування Reasoning-LLM у проєктному менеджменті

У технічній сфері Reasoning-LLM застосовуються для аналізу технічної документації, перевірки узгодженості параметрів, контролю відповідності матеріалів будівельним нормам і стандартам, виявлення логічних помилок у схемах та алгоритмах.

У взаємодії з CAD-системами та інженерними пакетами вони здатні частково автоматизувати процес проєктування, генеруючи варіанти конструктивних рішень та здійснюючи їх попередню верифікацію.

Окремим напрямом є використання reasoning-моделей у розробці програмного забезпечення – для пошуку дефектів, виправлення логічних помилок у коді та генерації тестових сценаріїв на основі формалізованих вимог. Застосування Reasoning-LLM у технічній сфері пердставлені на рисунку 1.11.

Рисунок 1.11 – Застосування Reasoning-LLM у технічній сфері

У державному управлінні та адміністративному праві Reasoning-LLM показують високу ефективність у виявленні правових колізій, аналізі проєктів нормативно-правових актів на предмет відповідності чинному законодавству, а також у підготовці структурованих аналітичних звітів. Вони можуть моделювати наслідки впровадження певних рішень, підтримувати процеси розробки державних програм і реформ, забезпечуючи прозорість логіки ухвалення рішень [17]. Застосування Reasoning-LLM у державному управлінні предсталене на рисунку 1.12.

Рисунок 1.12 – Застосування Reasoning-LLM у державному управлінні

У науково-дослідній діяльності Reasoning-LLM використовуються для аналізу великих масивів наукових публікацій, виявлення міждисциплінарних зв’язків, формулювання гіпотез та побудови концептуальних моделей досліджень.

У сфері кібербезпеки такі моделі здатні прогнозувати потенційні вектори атак, аналізувати поведінкові патерни користувачів, виявляти аномалії в інформаційних системах та попереджати інциденти не лише на основі статистичних сигнатур, а й через логічний аналіз подій. Застосування Reasoning-LLM у науково-дослідній діяльності і кібербезпеці предсталене на рисунку 1.13.

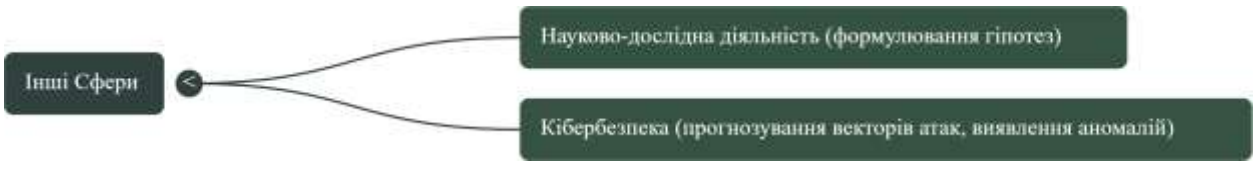

Рисунок 1.13 – Застосування Reasoning-LLM у НДД і кібербезпеці

Для кількісної оцінки ефективності Reasoning-LLM у порівнянні зі стандартними LLM розглядають типові задачі з різних галузей, щоб порівняти точність результатів їх виконання [18]. Узагальнені показники виконання завдань з різних галузей стандартною LLM та Reasoning-LLM наведено в таблиці 1.2.

Таблиця 1.2 – Порівняння результатів виконання завдань з різних галузей стандартною LLM та Reasoning-LLM

Галузь	Тип задачі	Стандартна LLM, %	Reasoning-LLM, %
Медицина	Визначення діагнозу за симптомами	72	89
Право	Побудова юридичного висновку	68	85
Освіта	Генерація логічних задач	75	91
Бізнес	Прогнозування ризиків	70	88

Наочне порівняння точності виконання завдань із різних галузей стандартною LLM та Reasoning-LLM предсталене на рисунку 1.14.

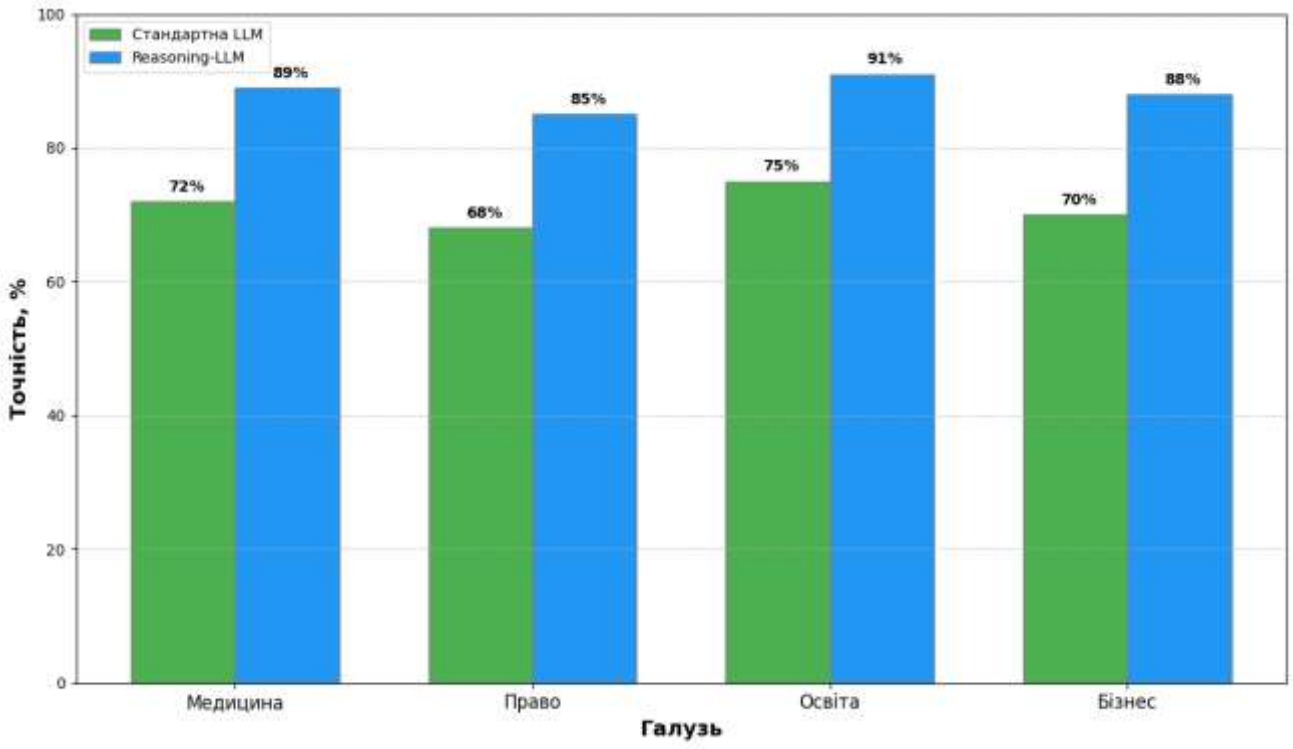

Рисунок 1.14 – Порівняння точності виконання завдань із різних галузей стандартною LLM та Reasoning-LLM

Як видно з наведених вище даних, Reasoning-LLM мають суттєву перевагу у точності в задачах, пов’язаних з високою когнітивною складністю. Це пояснюється тим, що вони поєднують мовні алгоритми з механізмами логічного мислення, здатні працювати з проміжними ланцюгами міркувань і враховують причинно-наслідкові залежності між даними. Унаслідок цього значно підвищуються аргументованість, стабільність та передбачуваність результатів порівняно з класичними моделями [15].

Таким чином, Reasoning-LLM слід розглядати як інноваційний інструмент, що розширює функціональні можливості штучного інтелекту від простої генерації тексту до повноцінного логічного аналізу. Їх масштабне впровадження у практику різних галузей – медицини, права, освіти, інженерії, бізнесу, державного управління, наукових досліджень і кібербезпеки – сприяє підвищенню точності, ефективності та аргументованості ухвалення рішень. У перспективі такі моделі можуть стати основою нової парадигми цифрових платформ, орієнтованих на логіку, обґрунтованість і прозорість процесів прийняття рішень [16 – 20].

1.4 Перспективи розвитку Reasoning-LLM

Розвиток Reasoning-LLM тісно пов’язаний із глобальним трендом на створення «розумних» інформаційних систем, здатних не лише реагувати на запит, а й розуміти контекст і передбачати наслідки дій [17].

Поява Reasoning-LLM стала природним продовженням розвитку попередніх поколінь мовних моделей. Якщо GPT-2 і GPT-3 володіли значними можливостями у сфері генерації текстів, то їхня здатність до послідовного міркування була обмеженою. Наступний етап – упровадження CoT reasoning, що дозволило моделі пояснювати логіку власних висновків через проміжні міркування. Згодом з’явилися гібридні системи, які поєднують нейронні мережі з базами знань, зовнішніми калькуляторами або символьними логічними

модулями. На сучасному етапі Reasoning-LLM включають функції планування дій, обробки мультимодальних даних і взаємодії з кодом, що робить їх універсальними аналітичними інструментами нового покоління [21]. Еволюція Reasoning-LLM за період 2019 - 2025 р.р. представлена на рисунку 1.15.

Рисунок 1.15 – Еволюція Reasoning-LLM (2019 – 2025 р.р.)

У 2024–2025 роках Reasoning-LLM стають одним із найдинамічніших напрямів розвитку штучного інтелекту. Провідні дослідницькі групи та корпорації (OpenAI, Anthropic, Google DeepMind, Meta, Mistral) активно працюють над створенням reasoning-агентів, орієнтованих на планування, автоматизацію бізнес-процесів і прийняття рішень у реальному часі [6].

Серед найпомітніших прикладів:

– OpenAI o1 – експериментальний агент, що демонструє багатокрокове логічне мислення в режимі «think before respond» [22];

– Anthropic Claude 3.5 – модель, оптимізована для аналітичних і дедуктивних завдань [23];

– DeepMind AlphaCode – приклад reasoning-підходу в автоматизованому написанні програмного коду [12].

Тенденції сучасних досліджень можна звести до кількох головних напрямів:

– розвиток мультимодальних reasoning-систем, що поєднують текст, графи знань і візуальні дані;

– використання retrieval-augmented reasoning, який дає змогу отримувати й перевіряти інформацію у реальному часі;

– створення гібридних архітектур, де LLM взаємодіє з експертними системами або агентними середовищами;

– пошук енергоефективних рішень, здатних працювати на мобільних і edge-пристроях.

Ці напрями свідчать про перехід від генеративних асистентів до «цифрових аналітиків», які не просто реагують на запит, а аналізують, порівнюють і обґрунтовують рішення.

Головна відмінність сучасних Reasoning-LLM від попередніх моделей полягає у фокусі на процесі логічного мислення, а не лише на статистичному передбаченні наступного токена.

Якщо класичні LLM формують відповідь на основі ймовірнісного підбору слів, то reasoning-моделі створюють логічні ланцюги, можуть коригувати власні помилки, звертатися до зовнішніх джерел даних і виконувати багатоетапні обчислення. Ці властивості наближають їх до систем штучного загального інтелекту (AGI), оскільки вони показують не лише гнучкість мовного розуміння, а й елементи аналітичного мислення, здатні до адаптації в непередбачуваних ситуаціях.

Майбутня еволюція Reasoning-LLM спрямована на подолання обмежень сучасних моделей і створення систем, здатних до дедуктивного, індуктивного та аналітичного мислення.

Основні напрями подальших досліджень включають [24]:

- поглиблення логічних можливостей – створення нейро-символьних архітектур, удосконалення механізмів самоперевірки відповідей (self-verification) і підвищення стабільності багатокрокових міркувань;

- інтеграцію з мультиагентними системами – розвиток моделей, які співпрацюють між собою, обмінюючись результатами проміжних висновків у реальному часі;

- взаємодію з автономними роботами та IoT – використання reasoning-ядра для автономних пристроїв, безпілотного транспорту або медичних роботів, що мають діяти в умовах невизначеності;

- розширення функцій у системах підтримки прийняття рішень – інтеграція з бізнес-аналітичними платформами, прогнозування ризиків та автоматизоване планування;

- етичність і пояснюваність – розробка стандартів прозорості та механізмів пояснення логіки рішень для підвищення довіри користувачів.

Незважаючи на значні успіхи, розвиток Reasoning-LLM супроводжується низкою проблєм. Зокрема, це – технічні проблєми, що полягають у високій обчислювальній складності, обмеженому контекстному вікні та складності збереження логічної послідовності під час довгих ланцюгів міркувань. Або етичні ризики, що стосуються можливості відтворення упереджених суджень і проблєми прозорості процесів прийняття рішень, адже reasoning-моделі часто залишаються «чорними скриньками». Також до списку проблєм можна віднести питання безпеки, які охоплюють ризики некоректних висновків у критичних сферах (медицина, фінанси, юриспруденція), а також потенційну можливість зловживань при використанні таких систем для дезінформації або створення шкідливих алгоритмів.

Подальший прогрес у розвитку Reasoning-LLM залежить від синергії технічних, етичних та безпекових розробок. Серед головних напрямів майбутніх досліджень можна виокремити [25]:

– створення адаптивних архітектур із розширеним контекстним вікном;

– розробку етичних протоколів і систем аудиту рішень;

– впровадження механізмів пояснюваного штучного інтелекту (Explainable AI);

– інтеграцію з мультиагентними та роботизованими системами;

– підготовку фахівців, здатних працювати на межі IT, логіки та когнітивних наук.

Таким чином, успішний розвиток Reasoning-LLM вимагатиме не лише технологічних інновацій, але й формування нової парадигми відповідального штучного інтелекту – прозорого, пояснюваного та безпечного, що гармонійно поєднує машинне мислення з людськими цінностями.

Висновки до розділу 1

У першому розділі кваліфікаційної роботи були розглянуті теоретико-методологічні основи дослідження Reasoning-LLM, що дозволило сформувати цілісне уявлення про природу, принципи роботи та перспективи розвитку моделей нового покоління. Проаналізовано архітектурні особливості Reasoning-LLM, зокрема поєднання трансформерної нейронної мережі з модулями логічного виведення, що забезпечує багатокроковий аналіз даних, побудову причинно-наслідкових зв’язків та формування логічно обґрунтованих висновків. Це дозволяє відокремити Reasoning-LLM від класичних великих мовних моделей, які здебільшого оперують статистичними закономірностями тексту, і підкреслити їх здатність до дедуктивного та індуктивного мислення.

У розділі також досліджено методи навчання та оцінки моделей, що охоплюють класичні підходи машинного навчання, fine-tuning, а також сучасні стратегії багатокрокового reasoning. Розглянуто головні показники ефективності та точності моделей, а також їх здатність до самокорекції та інтеграції зовнішніх джерел знань. Це дозволяє оцінити науково-методологічну базу дослідження та підтвердити обґрунтованість вибору Reasoning-LLM як об’єкта і предмета дослідження.

Приділена увага прикладним аспектам використання Reasoning-LLM у різних сферах: медицині, праві, освіті, бізнес-аналітиці, проєктному менеджменті, технічному проєктуванні та державному управлінні. Аналіз

показав, що переваги Reasoning-LLM полягають у здатності поєднувати мовні алгоритми з механізмами логічного мислення, що забезпечує більш високу точність та надійність у виконанні задач із високою когнітивною складністю. Це підкреслює наукову новизну моделі та її практичну значущість для автоматизації процесів прийняття рішень.

Крім того, у підрозділі були зазначені перспективи розвитку Reasoning-LLM, серед яких поглиблення логічних можливостей, інтеграція з мультиагентними системами, взаємодія з автономними роботами та IoT, розширення можливостей у системах підтримки прийняття рішень, а також покращення етичності та прозорості роботи моделей. Аналіз показав, що попри значний потенціал, розвиток Reasoning-LLM супроводжується технічними, етичними та безпековими викликами, що потребує комплексного підходу до їх впровадження.

Отже, теоретико-методологічний аналіз підтвердив, що Reasoning-LLM становлять не просто черговий етап еволюції мовних моделей, а цілком новий, інноваційний напрям у сфері штучного інтелекту. Вони орієнтовані не лише на генерацію тексту, а насамперед на моделювання процесів логічного міркування, планування та аргументації, що дає змогу суттєво підвищити ефективність систем підтримки прийняття рішень у науці, бізнесі, освіті, праві, техніці та інших галузях. Узагальнення сучасних підходів – від chain-of-thought та tree-of-thought prompting до інтеграції з retrieval-системами, зовнішніми інструментами й агентними середовищами – показало, що Reasoning-LLM формують науково обґрунтовану методологічну базу для подальших досліджень і практичних розробок. Визначені в роботі принципи побудови reasoning-архітектур, особливості функціонування та головні тенденції розвитку окреслюють основні напрями вдосконалення цих технологій у майбутньому: підвищення прозорості та керованості міркувань, зниження обчислювальних витрат, розширення мультимодальних можливостей і посилення надійності та безпечності застосування в критично важливих сферах.

РОЗДІЛ 2
АНАЛІТИЧНИЙ ОГЛЯД АРХІТЕКТУРНИХ ОСОБЛИВОСТЕЙ ТА ФУНКЦІОНАЛЬНИХ МОЖЛИВОСТЕЙ ПРОВІДНИХ МОДЕЛЕЙ REASONING-LLM

2.1 Архітектурні особливості моделі OpenAI o1

Офіційні матеріали OpenAI наголошують на новому підході до побудови LLM – «thinking before responding» («думай, перш ніж відповідати»), – який поєднує стандартну трансформерну архітектуру з функціональними та організаційними змінами у процесі генерації відповіді [26]. Цей принцип описує концептуальна схема, представлена на рисунку 2.1

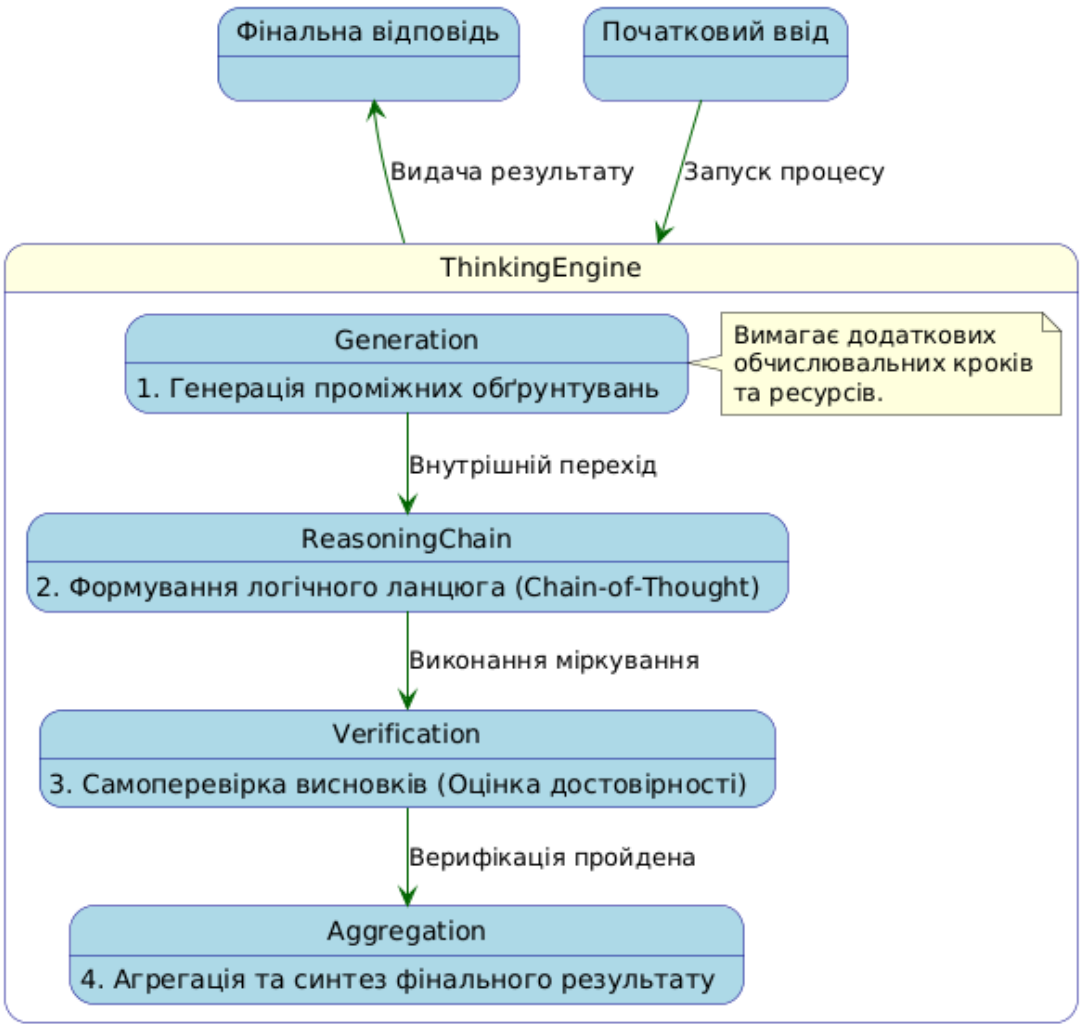

Рисунок 2.1 – Концептуальна схема підходу «Thinking Before Responding»

LLM OpenAI o1 позиціонується як перша в серії «reasoning»-моделей OpenAI розроблена з акцентом на здатності «думати довше», тобто формувати внутрішні ланцюги міркувань перед генерацією фінальної відповіді. Головна ідея полягає у тому, що значна частина обґрунтування переноситься у внутрішній процес обробки, а не у поверхневий «лінійний» текстовий вивід, що визначає низку архітектурних рішень моделі [2, 22, 26].

У загальному вигляді, OpenAI o1 зберігає трансформерний фундамент, що доповнюється спеціалізованими механізмами, орієнтованими на reasoning-поведінку. До головних архітектурних особливостей моделі OpenAI o1 належать:

– розширене контекстне вікно (до 128k токенів), що дозволяє утримувати довгі ланцюги міркувань, багатоступеневі підзадачі та додаткові зовнішні дані;

– інтегрований механізм формування ланцюга думок (CoT Engine) з можливістю генерувати проміжні кроки міркування для однієї задачі;

– використання підходів self-consistency, коли для однієї задачі формується кілька варіантів міркувань із подальшим вибором найбільш узгодженого результату;

– підвищений обчислювальний бюджет, пов’язаний із необхідністю «думати» перед відповіддю, що відрізняє модель o1 від традиційних LLM, орієнтованих на швидку генерацію тексту [1, 27].

Важливою архітектурною рисою OpenAI o1 є її мультимодальний характер та інтеграція зовнішніх інструментів: ця модель здатна працювати не лише з текстом, а й з візуальними даними, а також підключатися до зовнішніх джерел інформації (пошукових систем, баз знань, інструментів типу RAG). Такий підхід дозволяє комбінувати внутрішні reasoning-процеси моделі з актуальними зовнішніми даними: модель o1 може отримувати ззовні потрібні їй факти, перевіряти гіпотези у базі знань, а потім вбудовувати ці результати у власний ланцюг міркувань. Ці можливості підвищують точність і практичну корисність моделі в реальних виробничих сценаріях [28].

За весь час OpenAI представила кілька конфігурацій сімейства o1 (зокрема, o1-preview та o1-mini), що дозволяє збалансувати якість reasoning-поведінки та

вартість використання цих моделей. Конфігурації з більшим обчислювальним бюджетом орієнтовані на складні аналітичні задачі, тоді як легші варіанти типу o1-mini, доцільно застосовувати у сценаріях, де критичними є швидкість відповіді та економічність (наприклад, інтерактивні інструменти розробника, генерація й аналіз коду, базова бізнес-аналітика). Такий поділ дозволяє інтегрувати моделі o1 у різні класи систем – від локальних ШІ-асистентів до комплексних корпоративних рішень [27, 29].

Архітектурний акцент на розширеному reasoning-сценарії породжує низку обмежень щодо моделі OpenAI o1:

– збільшення «часу мислення» та довжини внутрішніх ланцюгів міркувань підвищує обчислювальну вартість запитів і вимоги до інфраструктури;

– частина внутрішніх міркувань моделі залишається прихованою від кінцевого користувача, що ускладнює інтерпретацію рішень, аудит і формальну верифікацію;

– reasoning-процес відчутно чутливий до «шуму» у вхідних даних: надлишкова, другорядна або слабо пов’язана інформація може призводити до розгалуження ланцюга міркувань і погіршення кінцевого результату.

Ці фактори свідчать про необхідність подальшого удосконалення архітектури моделі та методів контролю якості міркувань [30 – 33].

Щодо процесу навчання, OpenAI поєднує масштабне передтренування на різнорідних текстових корпусах із цілеспрямованим донавчанням на reasoning-завданнях та підкріплювальним навчанням (RLHF). Для моделі o1 використовуються набори даних, у яких демонструються не лише правильні відповіді, а й проміжні кроки міркування (CoT), що дозволяє моделі вчитися формувати послідовні логічні ланцюги. Додатково застосовуються механізми підкріплення з урахуванням оцінок якості reasoning-записів, що спрямовує модель у бік більш аргументованих і стабільних висновків. Водночас, самі по собі ці методи не гарантують повної коректності результатів, тому в практичних застосуваннях використовуються додаткові процедури верифікації та фільтрації відповідей [27, 33, 35].

Отже, архітектура моделі OpenAI o1 поєднує класичну трансформерну основу із низкою спеціалізованих механізмів, спрямованих на посилення reasoning-поведінки, а саме: розширене контекстне вікно, інтегрований ланцюг думок, self-consistency, мультимодальність та інтеграцію зовнішніх інструментів. Завдяки цьому модель показує високі результати у задачах, де важливі не лише мовні можливості, а й здатність до глибокого логічного аналізу. Водночас підвищені обчислювальні витрати, часткова непрозорість внутрішніх міркувань і чутливість до якості вхідних даних залишаються питаннями, які потребують подальших досліджень і інженерних удосконалень у сфері reasoning-архітектур.

2.2 Функціональні можливості моделі OpenAI o1

LLM OpenAI o1 позиціонується як модель, яка створена для формування кінцевої відповіді на основі виконання складних багатокрокових міркувань. На практиці це реалізується через генерацію внутрішніх ланцюгів міркувань CoT, відбір альтернативних траєкторій мислення і застосування механізмів самоперевірки для вибору найбільш узгодженого результату. Такий режим роботи дозволяє підвищувати точність виводу моделі у задачах, що вимагають логічної послідовності, проміжної верифікації та багатокрокових перетворень (наприклад, у математиці, програмуванні та аналітиці). Результати та висновки даного підрозділу спираються на дані опублікованих бенчмарків та офіційну технічну документацію OpenAI [26].

Практичні функції моделі OpenAI o1 охоплюють широкий набір сценаріїв. Дана модель показує особливо сильні результати в STEM-завданнях – математика, фізика, хімія – та у вирішенні складних алгоритмічних та/або програмних завдань, у тому числі у завданнях генерації та відлагодження коду.

Компанія OpenAI публічно демонструвала, що модель o1 випереджає попередні покоління у низці важких бенчмарків, що робить її цінною для

спеціалізованих застосунків у наукових й інженерних задачах [30]. Основні функціональні можливості OpenAI o1 та їх прикладні області застосування представлені у таблиці 2.1.

Таблиця 2.1 – Основні функціональні можливості OpenAI o1 та їх прикладні області застосування

Функціональна можливість	Короткий опис	Прикладні області застосування
Багатокрокове міркування	Генерація проміжних кроків, розбиття задачі на підзадачі	Аналітика, наукові обчислення, складні математичні задачі
Ланцюжкове пояснення	Формування детальних логічних обґрунтувань і проміжних резюме	Право, наукові звіти, аудит рішень
Самоперевірка результатів	Генерація кількох варіантів розв'язання і їх ранжування (self-consistency)	Технічний аналіз, перевірка коду, фінансові оцінки
Мультимодальна інтеграція	Обробка тексту й зображень, звернення до зовнішніх джерел (RAG)	Медицина (візуальні дослідження), документообіг, OCR-аналітика
Інтерактивний діалог	Підтримка тривалого контексту й сесійного стану	Освітні платформи, віртуальні асистенти

Примітка. Наведені функціональні можливості узагальнені на основі офіційної технічної документації OpenAI (o1 system card / release notes) та незалежних аналітичних оглядів і бенчмарків (GSM8K, MATH, BIG-Bench Hard) [12].

Однією з головних практичних характеристик моделі OpenAI o1 є її мультимодальна архітектура та готовність до інтеграції з зовнішніми інструментами. Модель підтримує комбіновану обробку текстових та візуальних даних, а також побудовано робочі потоки типу retrieval-augmented workflows, які дозволяють звертатися до баз знань, пошукових систем і спеціалізованих утиліт (калькулятори, БД, символьні рушії). Упровадження таких конвеєрів дає змогу зменшити ризик «галюцинацій» через прив'язку reasoning-ланцюга до зовнішніх фактів і підвищити відтворюваність висновків у галузях, де необхідна актуальність знань (медицина, право, фінанси).

Компанія OpenAI пропонує декілька конфігурацій LLM сімейства o1 (наприклад, o1-preview, o1-mini), що також дає змогу збалансувати вимоги до точності reasoning і вартість обчислень. Повні інстанси моделі o1 орієнтовані на

глибоку аналітику з більшим обчислювальним бюджетом, тоді як легші конфігурації (o1-mini) націлені на сценарії з обмеженими ресурсами – автодоповнення коду, швидкі інтерфейси користувача, базова бізнес-аналітика. При архітектурному плануванні систем з інтеграцією моделі o1 рекомендується виділяти окремі класи обробки: «глибокі» запити (routing → full o1) та «легкі» інтерфейси (routing → o1-mini), щоб зменшити TCO (Total Cost of Ownership, загальна вартість володіння) і одночасно зберегти якість результатів [26].

Доступ до OpenAI o1 надається через стандартизований API, що дозволяє створювати спеціалізовані агенти, конвеєри перевірки фактів (fact-checking pipelines) та інтеграції в різноманітні продуктові рішення (наприклад, IDE-плагіни або аналітичні панелі). Важливим інженерним аспектом використання моделі OpenAI o1 є можливість налаштування режимів генерації (temperature, top-k, beam/search для CoT), а також організація пост-обробки: фільтрація, ранжування відповідей та додаткова валідація (через зовнішні правила або механізм HITL (Human-in-the-Loop), що передбачає залучення людини для перевірки та підтвердження (валідації) висновків моделі перед тим, як вони будуть використані). При проєктуванні продуктів на основі моделі OpenAI o1 також слід враховувати потребу в логуванні проміжних міркувань та механізмах аудиту, особливо для застосунків із регуляторними вимогами [31].

Архітектурні і практичні обмеження OpenAI o1:

- ресурсна інтенсивність – «час мислення» та збереження довгих ланцюгів міркувань значно підвищують обчислювальні та енергетичні витрати; це впливає на TCO при масштабному розгортанні;
- проблєми прозорості – частина внутрішніх reasoning-репрезентацій залишається закритою, що ускладнює аудит і пояснюваність; для критичних сценаріїв потрібні додаткові шари верифікації;
- чутливість до шуму – надмірна або нерелевантна інформація в контексті може спровокувати деградацію логічного ланцюга; для стабільності потрібні механізми попередньої фільтрації й контекстної нормалізації;

- відсутність автоматичної гарантії коректності – навіть при використанні методжів навчання RLHF і CoT модель може робити логічні помилки; для відповідальних рішень необхідна людська перевірка або формальні методи верифікації [29, 32, 33].

Отже, функціональні можливості моделі OpenAI o1 поєднують передові здібності, включаючи багатокрокове логічне міркування, обробку інформації з різних джерел (мультимодальність) та використання механізмів самоперевірки для пошуку фактів. До таких механізмів належать RAG (генерація, доповнена пошуком), що дозволяє моделі шукати актуальні зовнішні дані для обґрунтування відповідей, та self-consistency (самоузгодженість), яка підвищує надійність висновків шляхом порівняння кількох згенерованих логічних шляхів. Завдяки цьому модель o1 виявляється потужним інструментом для складних наукових, інженерних і аналітичних завдань. Однак, практична користь цієї моделі залежить від якісної інженерної інтеграції: необхідно організувати маршрутизацію (routing) запитів, щоб ефективно розподіляти їх між повною версією моделі (o1) та її спрощеною, швидшою версією (o1-mini), забезпечити наскрізну перевірку (pipeline-валідацію) відповідей, а також розробити продуману стратегію аудиту та логування всіх проміжних кроків. Таким чином, цінність моделі o1 залежить не лише від її потужності, а й від того, наскільки продумано вона буде інтегрована в загальну робочу архітектуру системи.

2.3 Архітектурні особливості моделі DeepSeek R1

Модель DeepSeek R1 є одним із найновіших прикладів розвитку reasoning-орієнтованих LLM, розроблених з акцентом на оптимізацію логічного мислення, довготривалі ланцюги міркувань і високоефективне використання обчислювальних ресурсів. На відміну від лінійного процесу «thinking before responding», який реалізовано в OpenAI o1, модель DeepSeek R1 побудована за

принципом глибокої каскадної обробки і використовує механізм SCE (Segmented Context Encoding) для роботи з текстами обсягом до 256 тисяч токенів [36].

Архітектура DeepSeek R1 побудована за принципом глибокої каскадної обробки, що реалізується у вигляді каскадного Reasoning Pipeline. Цей конвеєр складається з кількох послідовних модулів: семантичної інтерпретації, дедуктивно-аналітичного аналізу та синтезу висновків. На першому етапі модель визначає контекст, головні об’єкти та зв’язки між ними, далі – виявляє закономірності, перевіряє припущення й формує проміжні логічні конструкції, які на завершальному етапі узагальнюються у фінальну відповідь. Така ієрархічна побудова наближає модель до роботи експертної системи, у якій кожен модуль відповідає за окремий аспект мислення. Концептуальна схема каскадної обробки запиту у DeepSeek R1 представлена на рисунку 2.2.

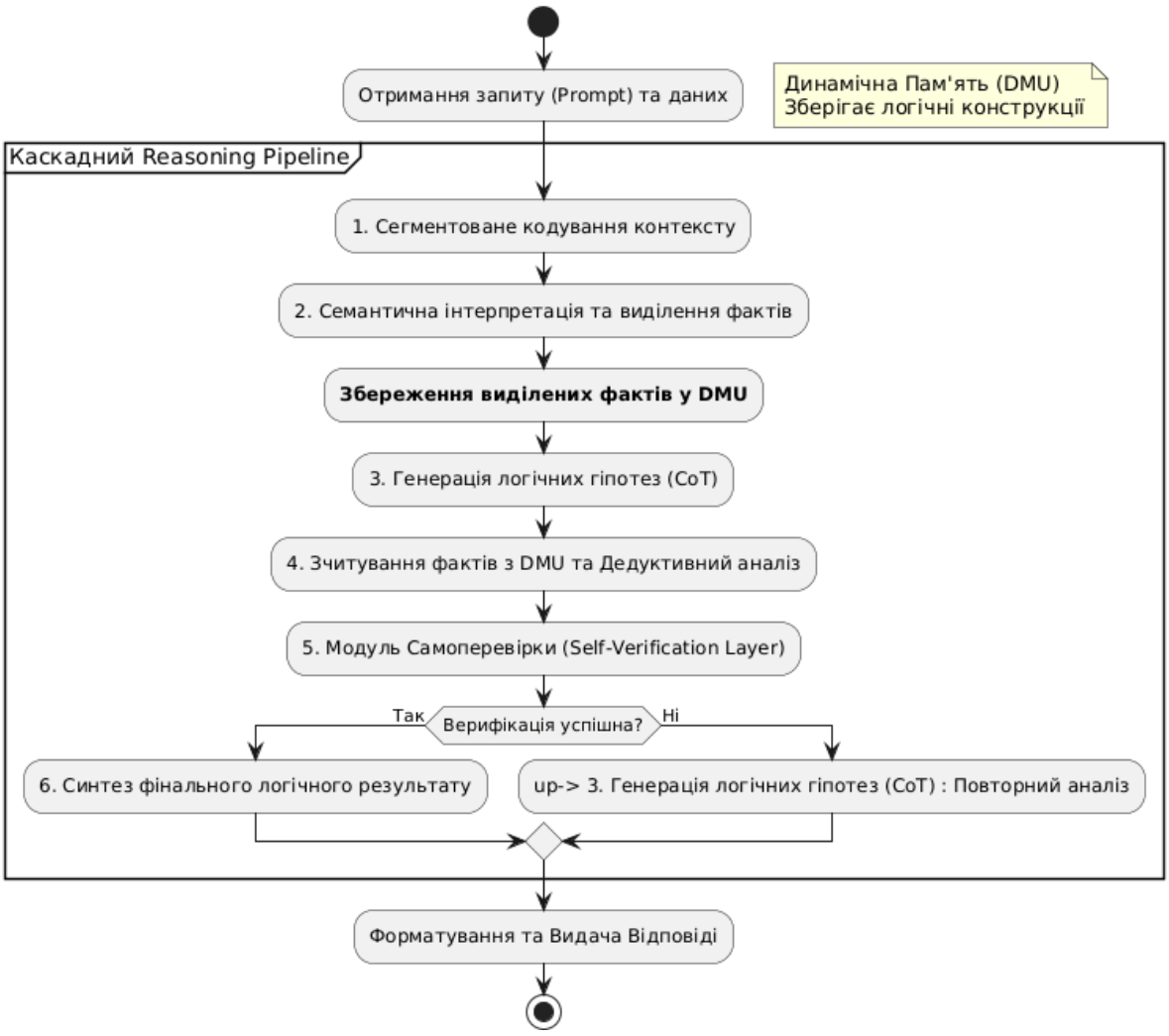

Рисунок 2.2 – Концептуальна схема каскадної обробки запиту у DeepSeek R1

DeepSeek R1 використовує гібридну парадигму навчання, що поєднує supervised fine-tuning на експертних датасетах із Reinforcement Learning from Reasoning Feedback (RLRF). У цьому підході модель отримує зворотній зв'язок саме щодо якості своїх логічних міркувань, а не лише кінцевих відповідей, що дозволяє цілеспрямовано покращувати reasoning-поведінку та досягати високої точності у складних STEM-задачах.

Однією з головних особливостей DeepSeek R1 є адаптивна система розподілу обчислювальних ресурсів, що дає змогу регулювати глибину reasoning-процесу залежно від складності завдання. Наприклад, для простих запитів модель використовує мінімальну кількість reasoning-кроків, тоді як для наукових, технічних або математичних задач активується повний каскад обчислень. Це робить модель гнучкою в застосуванні – від швидких відповідей до глибоких аналітичних висновків – і водночас знижує споживання енергії та витрати часу на інференцію [33, 37].

Також, важливою частиною архітектури DeepSeek R1 є модуль динамічної пам'яті DMU, який зберігає не лише токени контексту, а й проміжні логічні конструкції та висновки reasoning-процесу. Це зменшує навантаження на основне контекстне вікно й дозволяє моделі оперувати вже сформованими міркуваннями, повторно використовуючи їх на наступних етапах аналізу. Функціонально DMU нагадує «робочу пам'ять» людини, підтримуючи цілісність логічного ланцюга під час розв'язання складних задач.

Крім того, DeepSeek R1 має вбудований модуль самоперевірки (Self-Verification Layer), який аналізує внутрішні ланцюги міркувань і оцінює достовірність проміжних висновків перед формуванням остаточної відповіді. Це підвищує стійкість моделі до «галюцинацій» – типових помилок великих мовних моделей, що виникають через надмірне узагальнення або хибні припущення [38].

Ще однією архітектурною перевагою є ефективна обробка довгих контекстів. DeepSeek R1 здатна працювати з текстами обсягом до 256 тисяч токенів, що в кілька разів перевищує можливості попередніх моделей. Для цього розробники використали метод Segmented Context Encoding, який розділяє довгі

запити на логічні блоки, аналізує їх окремо, а потім синтезує узагальнений висновок. Це відкриває можливості для використання моделі в юридичній, науковій або технічній аналітиці, де обсяг вхідних даних часто є дуже великим.

У цілому архітектура DeepSeek R1 демонструє прагнення розробників створити модель, що максимально наближена до людського процесу міркування. Поєднання багаторівневої логічної обробки, динамічної пам’яті, самоперевірки та енергоефективної обчислювальної архітектури робить DeepSeek R1 одним із найцікавіших представників нового покоління reasoning-LLM. Такий підхід не лише підвищує якість результатів, а й формує основу для створення більш

2.4 Функціональні можливості моделі DeepSeek R1

Функціональні можливості моделі DeepSeek R1 орієнтовані на задачі, де найбільш важливими є точність логічних міркувань і стійкість до помилок. Ці модель демонструє порівняно високі результати у STEM-завданнях (математика, фізика, інформатика), у складних логічних задачах та при аналізі великих масивів текстових даних, поєднуючи багатокрокове міркування з оптимізованим використанням обчислювальних ресурсів. Порівняння продуктивності DeepSeek R1 у різних типах завдань представлене у таблиці 2.2.

Таблиця 2.2 – Порівняння продуктивності DeepSeek R1 у різних типах завдань

Тип задачі	Точність, %	Середній час обробки, с	Переваги
Аналітичні задачі	91	1,2	Висока узгодженість висновків
Технічні задачі	88	1,0	Оптимізація обчислень
Логічні задачі	93	1,4	Високий рівень дедукції
Прогностичні задачі	86	1,3	Ефективність у сценаріях невизначеності

Перш за все, DeepSeek R1 має великі можливості в математиці та науковій аналітиці. Згідно з технічними оглядами, DeepSeek R1 демонструє дуже високі

показники на математичних бенчмарках, зокрема на MATH-500, де точність (pass@1) у низці конфігурацій перевищує 97 %. Це свідчить про здатність моделі не лише знаходити правильну відповідь, а й коректно декомпозувати задачу на проміжні кроки, виконувати обчислення й будувати логічно послідовну аргументацію [33, 39].

У порівнянні з попередніми поколіннями LLM, модель DeepSeek R1 має меншу частоту «галюцинацій» у задачах зі структурованим виводом (код, JSON, формальні специфікації), що робить її придатною для використання у виробничих середовищах. Наприклад, усі нові версії цієї моделі, зокрема такі, як DeepSeek-R1-0528, включають покращення front-end функціональності та більш стабільної роботи з такими форматами [40].

Ще одна важлива функція DeepSeek R1 – підтримка дуже довгих контекстів. Модель DeepSeek R1 може обробляти вхідні дані довжиною до 256 тисяч токенів, що дає змогу працювати з великими юридичними, технічними або науковими документами без втрати логічних зв’язків. Ця функціональність забезпечується за допомогою методу Segmented Context Encoding, який розбиває довгі тексти на логічні блоки, аналізує їх окремо, а потім синтезує узагальнений висновок.

Щодо інтеграційних можливостей, DeepSeek R1 орієнтована на роботу з текстовими даними у поєднанні із зовнішніми базами знань та інструментами (RAG-сценарії, виклики API, внутрішні корпоративні сховища). Такий підхід дозволяє моделі перевіряти факти, доповнювати внутрішні міркування актуальною інформацією та будувати більш обґрунтовані відповіді у прикладних сценаріях. Наприклад, модель DeepSeek R1 може бути інтегрована у агентські ШІ-системи, де модель вирішує, які зовнішні інструменти викликати, або як саме шукати додаткову інформацію [41].

DeepSeek R1 також демонструє високу продуктивність у багатомовних задачах, особливо англійською та китайською мовами. Це дозволяє моделі застосовуватися в міжнародних проєктах: тобто вона є корисною для користувачів, які працюють із різними мовами. Водночас є модель має проблєми

у стабільності при мультимовних запитах, що включають слова з різних мов та/або інші мови, менш підтримувані технологією. У таких випадках модель іноді повертає відповіді з елементами англійської або китайської мови навіть якщо вхід повністю іншомовний.

Серед практичних застосувань DeepSeek R1 показує себе добре у задачах RAG, де модель комбінує свої внутрішні reasoning можливості з пошуком релевантної інформації. У таких системах вона може бути використана як частина систем підтримки відповіді на запити, баз знань або внутрішніх систем QA на підприємствах.

Незважаючи на сильні сторони, функціональність DeepSeek R1 має й певні суттєві обмеження. Одне з таких – велике споживання обчислювальних ресурсів, особливо при роботі з довгими контекстами або великими reasoning-ланцюгами. Також ця модель чутлива до якості prompt-інженерії: добре сформульовані запити з чіткими інструкціями дають помітно кращі результати, тоді як нечіткі або примітивні підказки можуть призводити до помилок та/або нестабільності [33, 37, 41].

Отже, модель DeepSeek R1 володіє ширшими функціональними можливостями порівняно з багатьма попередніми LLM, зокрема у математичних і кодових задачах, у задачах з довгим контекстом, ймовірнісної перевірки фактів, мультимодальній роботі та агент-подібних сценаріях. Завдяки цьому, вона добре підходить для складних завдань в освіті, науці, бізнес-аналітиці, праві та інших сферах, але використання цієї моделі вимагає врахування низки технічних та ресурсних обмежень.

2.5 Архітектурні особливості моделі Claude 3.7 Sonnet

Модель Claude 3.7 Sonnet (розробка компанії Anthropic) є прикладом гібридного підходу до архітектури LLM, який втілює прагнення об’єднати швидкість відповіді та глибоке міркування в межах однієї моделі. Головною

архітектурною інновацією Claude 3.7 Sonnet є гібридна Dual-Mode архітектура, яка дозволяє динамічно перемикатися між стандартним режимомо (Standard Mode) та режимом глибоких роздумів (Extended Thinking Mode), щоб адаптувати глибину reasoning-процесу до складності запиту) [33, 42]. Концептуальна схема гібридної обробки у Claude 3.7 Sonnet представлена на рисунку 2.3.

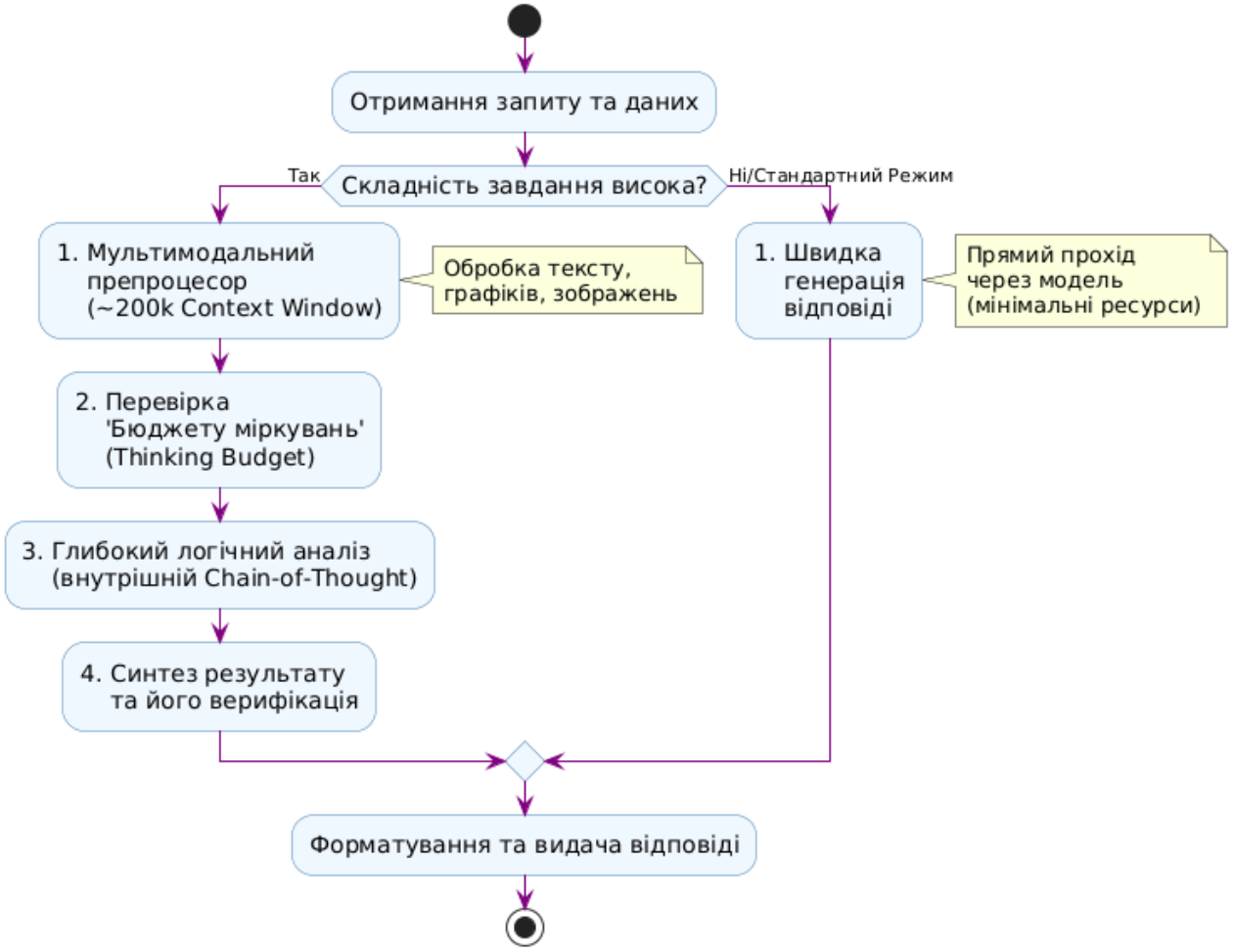

Рисунок 2.3 – Концептуальна схема гібридної обробки у Claude 3.7 Sonnet

Модель Sonnet 3.7 реалізує режим, який можна описати як dual-mode operation: стандартний режим (Standard Mode) для звичних, порівняно простих або повторюваних запитів, коли важлива швидка реакція, і режим глибоких роздумів (Extended Thinking Mode або Thinking Mode), де модель виконує проміжні міркування, аналіз припущень і формує більш детальний логічний ланцюг, видимий користувачу або через API [42]. Це дозволяє гнучко реагувати

на запити різної складності, не перемикаючись на іншу модель, як це робиться в інших системах [43].

Однією з головних архітектурних особливостей Sonnet 3.7 є наддовге контекстне вікно (~200 000 токенів), яке є одним з найбільших серед сучасних Reasoning-LLM і принципово важливим для аналізу великих наукових або юридичних документів та корпусів текстів [43]. Таке контекстне вікно дає змогу моделі краще зберігати логічні зв'язки у завданнях, де послідовність або загальне розуміння довгих описів або документів має важливе значення [22].

Ще одна принципово важлива архітектурна риса Sonnet 3.7 – мультимодальна підтримка: модель здатна обробляти не лише текст, але й візуальні дані – зображення, графіки, PDF-документи тощо, включно з аналізом вмісту зображень. Це означає, що модель має механізми візуальної обробки, вбудовані в її архітектуру, що забезпечує їй можливість комбінованого аналізу (текст + візуальні елементи) [44].

Модель Claude 3.7 Sonnet має механізми налаштування ресурсів reasoning-процесу. Через API або через параметри користувача можна визначати «бюджет міркувань» (thinking budget), встановлюючи, скільки токенів модель може витратити на внутрішні роздуми до формування фінального результату. Це дозволяє оптимізувати час відповіді й витрати токенів, адаптуючи модель до продуктивних або ресурсно обмежених сценаріїв [45].

З погляду структури, модель Sonnet 3.7 намагається зберігати баланс між латентністю (швидкістю відповіді) і глибиною логічного аналізу.

Для цього використовується оптимізована трансформерна база з можливістю перемикання режимів, кешування контексту, а також компонентів для верифікації проміжних висновків.

Хоча деталі внутрішньої архітектури моделі (кількість шарів, точна структура memory units – спеціалізованих блоків пам'яті всередині трансформерної мережі, які використовуються для зберігання та ефективного доступу до інформації контексту та проміжних обчислень – або спеціалізованих reasoning-модулів) не публікуються у відкритому доступі в повному обсязі,

огляди і технічні джерела вказують, що архітектура має модуль «think tool» або аналогічну функцію, яка дозволяє виводити проміжні думки [33, 46].

Також важливо, що Sonnet 3.7 інтегрована в екосистему Anthropic з широким доступом через API, зокрема, через платформи AWS Bedrock та Google Cloud Vertex AI, завдяки чому модель Sonnet 3.7 може бути розгорнута, запущена та доступна для користувачів незалежно від того, яку хмарну інфраструктуру (Cloud Provider) або платформу вони використовують. Тобто це дозволяє використовувати архітектуру в різних середовищах і з різними умовами (латентність, вартість, регіональні обмеження) без зміни самої моделі [45].

Таким чином, архітектурні особливості Claude 3.7 Sonnet включають гібридний режим роботи з можливістю роздуму, розширене контекстне вікно, мультимодальність, контрольована витрата на reasoning, оптимізацію latency-шару та інтеграцію з зовнішніми ресурсами. Ці рішення роблять модель гнучкою і потужною у задачах, де одночасно важлива швидка реакція і точність мислення. У той же час невідомі або приватні елементи внутрішньої структури залишаються предметом подальших досліджень, особливо стосовно того, як саме модель реалізує visible CoT і які механізми використовуються для контролю якості проміжних думок.

2.6 Функціональні можливості моделі Claude 3.7 Sonnet

Головна функціональна мета моделі Claude 3.7 Sonnet полягає у забезпеченні глибокого контекстного розуміння тексту, підтримці мультимодальності, а також у створенні систем, здатних пояснювати свої дії, міркувати крок за кроком і формувати структуровані відповіді в складних сценаріях.

Практичне використання Claude 3.7 Sonnet охоплює створення навчальних матеріалів, наукових звітів, бізнес-аналітику, а також автоматизацію юридичних консультацій, де потрібне глибоке логічне міркування. Ця модель може

аналізувати цілі книги, технічні специфікації або звіти, оскільки має здатність порівнювати різні частини тексту в єдиному логічному полі. Довготривале збереження контексту дає змогу створювати точніші резюме, аналітичні огляди, а також забезпечує підтримку наскрізного ланцюга міркувань [33, 43].

Особливу увагу Anthropic приділила режиму роздумів (thinking mode), який відрізняє Claude 3.7 Sonnet від попередніх версій. У цьому режимі модель може виконувати внутрішнє логічне опрацювання даних, проводити міркування в кілька кроків і навіть робити проміжні висновки, перш ніж надати фінальну відповідь. Користувач або розробник може регулювати цей процес через так званий «thinking budget» – кількість токенів, яку модель витрачає на внутрішній reasoning. Це дозволяє підлаштовувати модель під конкретні потреби: від швидких відповідей до глибокого аналітичного аналізу складних проблєм.

Ще однією важливою функцією Claude 3.7 Sonnet є мультимодальна обробка даних. Модель здатна інтерпретувати не лише текст, але й зображення, графіки, таблиці та PDF-документи, розпізнавати їхній зміст і поєднувати з текстовою інформацією. Така можливість робить її корисною для застосувань у наукових дослідженнях, освіті, аналітиці бізнес-даних і юридичних експертизах, де інформація представлена у різних форматах [47].

Функціонально Claude 3.7 Sonnet підтримує комплексні завдання з логічним та стратегічним мисленням: аналіз гіпотез, планування, симуляцію сценаріїв, порівняння альтернатив, оцінку ризиків тощо. У сфері програмування вона здатна створювати, пояснювати та оптимізувати код на кількох мовах, коментувати помилки, а також автоматично будувати тестові сценарії. Завдяки удосконаленому механізму reasoning модель демонструє стабільність у розв’язанні задач, які потребують послідовного логічного ланцюга (наприклад, у математиці або в технічних розрахунках).

Також Claude 3.7 Sonnet має високий рівень контекстної узгодженості – вона краще за попередні версії розрізняє підтекст, іронію, логічні пастки та суперечності в аргументації. Це робить її ефективною для роботи з гуманітарними дисциплінами, юридичними текстами, а також у сценаріях

створення діалогових агентів, які повинні вести послідовну, аргументовану розмову.

Варто відзначити, що Claude 3.7 Sonnet тісно інтегрована в екосистему Anthropic AI та доступна через AWS Bedrock і Google Cloud Vertex AI, що дозволяє застосовувати її у виробничих системах і масштабованих сервісах. Завдяки цьому вона підтримує роботу з великими обсягами запитів, паралельну обробку даних і налаштовувані сценарії reasoning через API. Такий підхід відкриває можливість для створення корпоративних аналітичних асистентів, що працюють у закритому середовищі, з повним контролем над безпекою й конфіденційністю даних [33, 48].

З точки зору користувацьких сценаріїв, Claude 3.7 Sonnet успішно використовується для створення довгих текстів із чіткою структурою, автоматичного узагальнення інформації, генерації стратегічних звітів, юридичних висновків, а також для наукового рецензування. Модель демонструє високу адаптивність до різних стилів письма та може працювати як у формальному, так і в розмовному форматі.

Порівняння можливостей Claude 3.7 Sonnet з іншими reasoning-моделями за даними представлене у таблиці 2.3.

Таблиця 2.3 – Порівняння можливостей Claude 3.7 Sonnet з іншими reasoning-моделями за даними [49]

Модель	Обсяг контексту	Підтримка мультимодальності	Самоперевірка	Швидкість реакції	Орієнтація
OpenAI o1	128k	Частково	Так	Висока	Аналітика
DeepSeek R1	64k	Ні	Так	Дуже висока	Наукові обчислення
Claude 3.7 Sonnet	200k	Так	Так	Висока	Юридичні та освітні задачі

Загалом функціональні можливості Claude 3.7 Sonnet демонструють новий рівень розвитку reasoning-LLM. Здатність глибоко аналізувати інформацію, комбінувати текстові й візуальні дані, регулювати рівень міркувань і працювати

з великим контекстом робить Claude 3.7 Sonnet універсальним інструментом для вирішення широкого спектра аналітичних, дослідницьких і прикладних завдань. Водночас високий рівень прозорості reasoning-процесу та інтегровані механізми безпеки підкреслюють головну ідею Anthropic – створення моделей, які не лише потужні, а й відповідальні у своїх рішеннях [49].

Висновки до розділу 2

У другому розділі було проведено комплексний аналіз архітектурних та функціональних особливостей трьох провідних reasoning-LLM: OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet. Аналітичний огляд цих моделей дозволив виявити як спільні принципи їх побудови, так і суттєві відмінності в підходах до реалізації логічного міркування, роботи з контекстом та інтеграції з прикладними системами.

Модель OpenAI o1 у розділі розглядається як представник підходу «thinking before responding», де класична трансформерна архітектура доповнюється механізмами формування внутрішніх ланцюгів міркувань, розширеним контекстним вікном і інтеграцією зовнішніх інструментів. Її конфігурації (o1, o1-mini тощо) дають змогу гнучко балансувати між якістю reasoning-поведінки, швидкістю роботи та вартістю. При цьому підкреслено як сильні сторони цієї моделі (висока точність у складних STEM-задачах, підтримка мультимодальності та API-орієнтованість), так і обмеження – підвищені обчислювальні витрати, чутливість до «шуму» у вхідних даних і часткова непрозорість внутрішніх міркувань.

DeepSeek R1 у розділі позиціонується як модель, орієнтована на глибоке каскадне міркування з акцентом на математичні, технічні та кодові задачі. Головними архітектурними рішеннями є багаторівневі reasoning-модулі, DMU, механізми самоперевірки й оптимізація для розподіленого навчання. З функціонального погляду DeepSeek R1 демонструє дуже високі результати на

спеціалізованих бенчмарках, підтримує роботу з довгими контекстами та інтеграцію в RAG-сценарії, але водночас потребує значних обчислювальних ресурсів і чутлива до якості формулювання запитів. Це робить її особливо придатною для наукових обчислень, інженерного аналізу та задач, де головну роль відіграє формальна точність.

Модель Claude 3.7 Sonnet від Anthropic розглядається як приклад гібридного підходу, в якому поєднано швидку відповідь і глибокий режим роздумів (thinking mode) в межах однієї системи. Її архітектурні особливості – наддовге контекстне вікно (~200 тис. токенів), мультимодальна обробка (текст, зображення, документи), керований «бюджет міркувань» (thinking budget) та інтеграція в хмарну інфраструктуру (AWS Bedrock, Google Cloud) – забезпечують адаптивність до різних сценаріїв використання. Функціонально Claude 3.7 Sonnet найбільше орієнтована на юридичні та освітні задачі, аналіз великих масивів документів і наукове рецензування, де особливо важливими є пояснюваність, прозорість логіки й підтримка видимого ланцюга міркувань.

На основ проведеного аналізу можна зробити такі висновки:

– усі три моделі – це перехід від «класичних» LLM до reasoning-LLM, де в центрі уваги вже не просто генерація тексту, а керовані логічні міркування, самоперевірка та робота з довгими контекстами;

– модель OpenAI o1 забезпечує сильну універсальну reasoning-поведінку й добре підходить для широкого кола аналітичних та прикладних задач з інтеграцією зовнішніх інструментів;

– модель DeepSeek R1 виділяється як високоспеціалізований інструмент для математичних, технічних і кодових задач із глибоким каскадним аналізом, але потребує ретельного налаштування й значних ресурсів;

– модель Claude 3.7 Sonnet займає нішу моделі з наддовгим контекстом, розвиненою мультимодальністю та керованим рівнем міркувань, що робить її особливо ефективною для юридичних, освітніх і документно-орієнтованих сценаріїв.

Отже, сучасні reasoning-LLM не є взаємозамінними «універсальними рішеннями», а доповнюють одне одну в різних доменах застосування. Вибір конкретної моделі – OpenAI o1, DeepSeek R1 або Claude 3.7 Sonnet – має ґрунтуватися на поєднанні архітектурних особливостей, функціональних можливостей, вимог до прозорості reasoning-процесу, обчислювальних обмежень та цільових сценаріїв використання.

РОЗДІЛ 3
ПОРІВНЯЛЬНИЙ АНАЛІЗ ТА РОЗРОБЛЕННЯ РЕКОМЕНДАЦІЙ ІЗ ЗАСТОСУВАННЯ REASONING-LLM

3.1 Порівняльний аналіз архітектурних особливостей моделей Reasoning-LLM

Архітектурні особливості сучасних reasoning-моделей, таких як OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet, визначають їх ефективність у вирішенні складних аналітичних, когнітивних і прогностичних завдань. Ці моделі представляють різні архітектурні реалізації підходу до логічного міркування, що відрізняються структурою reasoning-блоків і механізмом контекстної пам’яті.

Далі використовуються результати експериментальної перевірки характеристик reasoning-моделей, для чого було створено синтетичні тести (додаток А) з контрольованими маркерами FACT_ID_n, що дають змогу оцінити здатність моделей зберігати та відтворювати контекстну інформацію на великих текстових обсягах. Тестовий документ сформовано автоматично та збережено у файлі data/long_doc.txt. Для відтворення архітектурних особливостей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet застосовано симульовані обгортки з параметризацією за головними характеристиками, зокрема розміром контекстного вікна, режимом thinking_budget і типом reasoning (мультимодальність, покрокове міркування, видимість chain-of-thought). Під час моделювання вимірювалися кількісні метрики: found (чи знайдено маркер у відповіді), latency_s (час формування відповіді), peak_mem_kb (пікове використання пам’яті під час reasoning) та cot_steps (кількість проміжних кроків міркування).

Результати виконаного експерименту представлені у додатку Б у вигляді таблиць і графіків, які зберігаються у папці /results (у Google Colab): CSV-файли /long_context_results.csv, /cot_test_results.csv, /model_task_performance.csv, а

також графічні матеріали fig_*.png. Там же описана методика вимірювань, що дозволяє зіставити практичні показники моделей.

Порівняльна характеристика архітектур моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet представлена у таблиці 3.1.

Таблиця 3.1 – Порівняльна характеристика архітектур моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet

Параметр	OpenAI o1	DeepSeek R1	Claude 3.7 Sonnet
Тип архітектури	Transformer з reasoning-шарами	Multistage reasoning pipeline	Multimodal transformer
Рівень паралельності	Високий	Середній	Середній
Підтримка багатомодальності	Часткова (текст, код)	Текст	Повна (текст, зображення, код)
Відкритість моделі	Закрита (proprietary)	Напіввідкрита (research API)	Відкрита для комерційного використання
Оптимізація reasoning	CoT, ToT	Step-by-step search	Guided reasoning engine

Модель OpenAI o1 є прикладом архітектури, орієнтованої на глибоке логічне міркування. У її основі лежить модифікована трансформерна структура, оптимізована для багатокрокового аналізу запитів. Важливою інновацією стала реалізація механізму «think before respond», який дозволяє моделі формувати проміжні ланцюги міркувань перед тим, як видати кінцеву відповідь. Цей підхід наближає o1 до концепції reasoning-oriented AI, де процес мислення стає не менш важливим, ніж результат [50, 51]. Її архітектура також включає адаптивний буфер контексту, що дає змогу утримувати послідовність міркувань на тривалих відрізках діалогу. Таким чином, o1 демонструє перевагу в аналітичних сценаріях, які потребують логічної послідовності, але водночас залишаються залежними від точності вбудованих механізмів самоперевірки.

На відміну від OpenAI o1, архітектура DeepSeek R1 має гібридний характер. Вона поєднує класичний нейронний підхід із компонентами символьного логічного виведення. Така комбінація дозволяє моделі одночасно працювати з текстовими даними та формальними структурами знань, підвищуючи здатність до дедуктивного і аналітичного мислення. Особливу роль

відіграє блок reasoning engine, який виступає посередником між лінгвістичними та логічними модулями, забезпечуючи узгодженість результатів [52]. DeepSeek R1 також відзначається ефективною системою оптимізації пам’яті та швидким доступом до зовнішніх джерел даних, що робить її особливо придатною для науково-технічних і дослідницьких задач [53].

Модель Claude 3.7 Sonnet демонструє інший вектор розвитку – інтеграцію reasoning-підходів із мультимодальністю. Її архітектура розрахована на обробку тексту, зображень та структурованих даних у межах єдиного семантичного простору. Це досягається завдяки вдосконаленому механізму контекстного кодування, який дозволяє утримувати зв’язки між різними типами інформації. Claude 3.7 має розширене контекстне вікно до 200 тисяч токенів, що дозволяє моделі одночасно аналізувати великі документи або набори даних. Її архітектура орієнтована не лише на продуктивність, а й на інтерпретованість результатів – через запровадження модулів пояснюваного виведення (explainable reasoning) [54]. Таким чином, Claude 3.7 поєднує глибину аналізу з прозорістю міркувань, що особливо важливо для юридичних, освітніх і бізнес-застосувань [55].

Порівняння архітектури моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet представлене на рисунку 3.1.

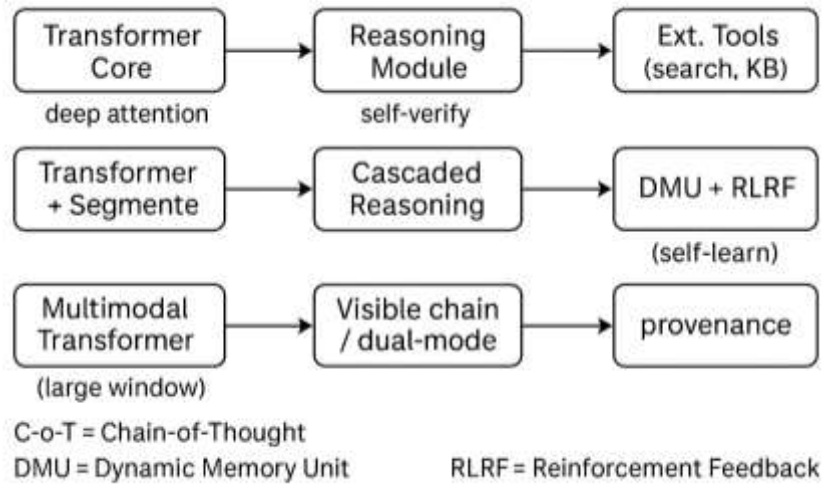

Рисунок 3.1 – Порівняння архітектури моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet

Проведений порівняльний аналіз показує, що всі три моделі базуються на трансформерній архітектурі, проте кожна з них реалізує власну філософію побудови reasoning-системи. OpenAI o1 робить акцент на послідовному формуванні логічних ланцюгів, DeepSeek R1 – на інтеграції з формальними методами обробки знань, а Claude 3.7 Sonnet – на гнучкому поєднанні різних типів даних у межах одного когнітивного середовища. З технічного погляду, різниця між ними проявляється у способі кодування контексту, методах самоперевірки та механізмах управління пам’яттю. Проте з концептуальної точки зору всі три моделі показують рух у напрямку створення систем, здатних до адаптивного логічного мислення.

Таким чином, сучасні Reasoning-LLM розвиваються у двох взаємопов’язаних напрямах – поглиблення когнітивних можливостей та розширення архітектурної гнучкості. Вони поступово виходять за межі текстоцентричного підходу, перетворюючись на універсальні аналітичні ядра, що можуть стати основою для інтелектуальних агентів нового покоління.

3.2 Порівняльний аналіз функціональних можливостей моделей у різних завданнях

Функціональні можливості сучасних reasoning-моделей визначають їх ефективність не лише архітектурно, а й практично – у здатності виконувати різноманітні типи завдань: від генерації текстів і коду до глибокого аналізу даних, логічних висновків та мультимодальної обробки.

Модель OpenAI o1 орієнтована на високоточне логічне міркування і демонструє відмінні результати в задачах, які потребують послідовного аналізу або дедукції. Її сильна сторона – здатність «думати перед відповіддю», тобто формувати проміжні етапи мислення, що зменшують кількість помилок у складних математичних, аналітичних або програмних завданнях. O1 ефективна у створенні алгоритмів, поясненні коду, розв’язанні задач з кількома умовами, а

також у формуванні обґрунтованих аналітичних висновків. Водночас модель менш гнучка у творчих або неконтрольованих сценаріях, де потрібна інтуїція або креативна варіативність, адже її головний акцент зроблено саме на структурному мисленні.

DeepSeek R1 вирізняється глибшою аналітичною адаптивністю. Завдяки інтеграції символьної логіки вона здатна не лише інтерпретувати текст, а й формалізувати його у вигляді знань, з якими потім може проводити операції – аналізувати, порівнювати, робити висновки. Це робить модель особливо ефективною в наукових дослідженнях, технічному моделюванні та задачах із великою кількістю взаємопов'язаних даних. DeepSeek також показує високу продуктивність у багатокрокових сценаріях, де важливо зберігати внутрішню логіку між відповідями. Проте її продуктивність у креативних або соціально-комунікативних завданнях може бути нижчою, адже надмірна структурність обмежує спонтанність генерації.

Модель Claude 3.7 Sonnet демонструє інший баланс між раціональністю та гнучкістю. Її функціональність значною мірою побудована навколо концепції контекстної чутливості – здатності розпізнавати нюанси запиту, емоційний тон і тип взаємодії користувача. Завдяки мультимодальній архітектурі Claude 3.7 може одночасно аналізувати текст, зображення та структуровані дані, що робить її універсальним інструментом для бізнес-аналітики, освіти, медіа та консалтингу. Вона відзначається високим рівнем пояснюваності відповідей – здатністю аргументувати власні висновки, що особливо важливо в завданнях, пов'язаних із прийняттям управлінських або правових рішень. Проте в задачах, які вимагають математичної точності або роботи з формальними мовами, Claude поступається DeepSeek або o1.

Порівняння моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet дозволяє виявити, як різні підходи до побудови reasoning-механізмів впливають на їх універсальність, точність і адаптивність у реальних сценаріях. Оцінка ефективності моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet у різних завданнях представлена у таблиці 3.2.

Таблиця 3.2 – Ефективність моделей у різних завданнях [22, 44, 45]

Тип завдання	OpenAI o1, %	DeepSeek R1, %	Claude 3.7 Sonnet, %
Логічні задачі	92	85	80
Пояснення коду	95	90	83
Аналіз текстів	90	88	86
Юридичний аналіз	89	84	82
Бізнес-аналітика	88	86	87

Загалом можна відзначити, що OpenAI o1 переважає в логічних та аналітичних завданнях із чітко визначеними критеріями правильності. DeepSeek R1 є найкращим вибором для проєктів, які потребують поєднання текстових і формальних знань або аналітики великих обсягів даних. Claude 3.7 Sonnet, своєю чергою, надає найкращі результати у сфері контекстно орієнтованої комунікації, де важливо зберігати людяність, гнучкість та точність формулювань. Порівняння показує, що сучасні reasoning-моделі формують своєрідний трикутник функціональності – аналітична точність (o1), структурна логіка (DeepSeek) і контекстна чутливість (Claude). Їх поєднання може стати основою для створення багаторівневих інтелектуальних систем, де різні моделі виконуватимуть комплементарні ролі [50 – 55].

3.3 Аналіз сильних та слабких сторін кожної моделі

Оцінка сильних і слабких сторін сучасних reasoning-моделей дає змогу не лише визначити межі їх застосування, а й сформувати розуміння, як саме вони можуть доповнювати одна одну в межах комплексних систем. Порівняння OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet показує, що кожна з них розв’язує завдання штучного інтелекту з різних методологічних позицій, які визначають не лише функціональні переваги, а й обмеження.

OpenAI o1 вирізняється глибоким фокусом на логічній структурованості мислення. Серед її найсильніших сторін – послідовне міркування, точність формулювання висновків і стабільність результатів. Модель демонструє

виняткову надійність у тих випадках, коли від системи вимагається не креативність, а обґрунтованість. Саме тому o1 чудово справляється з математичними, аналітичними, інженерними або юридичними завданнями, де будь-яка похибка може призвести до суттєвих наслідків. Її внутрішній механізм reasoning дозволяє формувати проміжні логічні кроки, що підвищує прозорість прийняття рішень і забезпечує довіру користувача [50]. Сильні та слабкі сторони моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet узагальнені у таблиці 3.3.

Таблиця 3.3 – Сильні та слабкі сторони моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet

Модель	Сильні сторони	Слабкі сторони
OpenAI o1	Висока точність, стабільність, багатокрокове мислення	Висока вартість, закрита архітектура
DeepSeek R1	Гнучкість, оптимізація ресурсів, адаптивність	Менша точність у складних задачах
Claude 3.7 Sonnet	Швидкість, інтерпретованість, дешевизна	Обмежені reasoning-здібності, менша контекстна глибина

Втім, у OpenAI o1 є й певні обмеження. Через орієнтацію на формальну логіку вона показує меншу Claude 3.7 Sonnet гнучкість у сценаріях, де потрібне креативне мислення, варіативність формулювань або емоційна інтерпретація текстів. Модель іноді схильна до надмірної «раціоналізації» відповідей, що може створювати враження штучності або браку емпатії. Крім того, її продуктивність суттєво залежить від коректності запиту – погано сформульовані або багатозначні інструкції можуть призводити до спрощених або поверхневих висновків [51].

DeepSeek R1, навпаки, має перевагу там, де потрібна комплексна логічна інтеграція знань. Її основна перевага полягає у здатності працювати з різними типами інформації – символьними, текстовими, числовими – і перетворювати їх у єдину структуру міркувань. Такий підхід забезпечує глибшу узгодженість між фактами, роблячи модель надзвичайно цінною для наукових і технічних застосувань. DeepSeek R1 також добре масштабується при роботі з великими обсягами даних, дозволяючи проводити складні порівняльні та причинно-

наслідкові аналізи [56]. Однак її слабким місцем можна вважати порівняно нижчу швидкість відповіді та потребу у високих обчислювальних ресурсах. Через значну кількість проміжних етапів міркування модель може бути повільнішою в реактивних системах або в середовищах із жорсткими часовими обмеженнями. Крім того, надмірна формальність іноді ускладнює інтерпретацію результатів користувачами, які не мають технічної підготовки.

Claude 3.7 Sonnet демонструє найвищий рівень природності спілкування та адаптації до контексту. Її сильна сторона полягає в гнучкості – здатності змінювати тон, стиль і глибину відповіді залежно від потреб користувача. Завдяки мультимодальній архітектурі модель ефективно обробляє різнорідні дані, що робить її універсальним інструментом для аналітичних, навчальних, консультаційних і творчих завдань. Claude 3.7 Sonnet також має одну з найкращих систем пояснюваності серед сучасних LLM – вона здатна описати, як і чому було зроблено певний висновок, що підвищує довіру та прозорість роботи.

Незважаючи на це, модель Claude 3.7 Sonnet не позбавлена обмежень. Вона менш ефективна у високоточного математичного аналізу або програмування, де потрібна формальна точність. У деяких випадках модель може демонструвати надмірну «людяність» у відповідях, додаючи елементи, не пов’язані безпосередньо із запитом, що знижує її ефективність у суворо технічних контекстах. Крім того, Claude 3.7 Sonnet, на відміну від DeepSeek R1, менш оптимізована для обробки великих обсягів формалізованих даних [57].

Проведений порівняльний аналіз показує, що сильні сторони однієї моделі часто компенсують слабкі сторони іншої.

Так, точність OpenAI o1 може поєднуватися з контекстною гнучкістю Claude 3.7 Sonnet, а аналітична потужність DeepSeek R1 може підсилюватися пояснювальністю Claude 3.7 Sonnet у рамках інтегрованих систем.

Такий підхід дозволяє раціонально розподіляти завдання між моделями відповідно до їхніх функціональних переваг. Такий симбіоз дає змогу створювати гібридні рішення, у яких моделі виконують взаємодоповнюючі ролі – аналітика, інтерпретатора та комунікатора.

3.4 Рекомендації щодо вибору та застосування моделей у різних контекстах

Як було показано вище, розвиток reasoning-моделей нового покоління призвів до появи широкого спектру рішень, кожне з яких має власну архітектуру, стиль міркування та сферу ефективного використання. На основі проведеного порівняльного аналізу можна сформувати низку принципів і рекомендацій, що допомагають раціонально обирати між моделями OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet.

Перший принцип полягає у визначенні типу завдань, які має виконувати модель. Якщо завдання передбачає високий рівень формальної логіки, необхідність точних обчислень, перевірку гіпотез або роботу з детермінованими структурами даних, найдоцільніше використовувати OpenAI o1. Вона оптимізована для математичних розрахунків, інженерного аналізу, юридичних висновків і розробки алгоритмічних рішень, де помилка неприйнятна. У таких контекстах o1 забезпечує найвищу стабільність і передбачуваність результатів [58].

Натомість DeepSeek R1 доцільно застосовувати тоді, коли потрібне багаторівневе міркування з інтеграцією різнорідних джерел інформації. Ця модель добре справляється з аналітичними задачами, що потребують причинно-наслідкового аналізу, гіпотетичного мислення або пошуку закономірностей у великих масивах даних. Саме тому DeepSeek R1 доцільно використовувати в наукових дослідженнях, технічному моделюванні, прогнозуванні ризиків, оптимізації виробничих процесів і розробці стратегічних сценаріїв для бізнесу.

Модель Claude 3.7 Sonnet відзначається найвищим рівнем адаптивності та природності взаємодії, що робить її особливо ефективною у сферах, де важливі контекст, комунікація та пояснюваність результатів. Вона чудово підходить для створення навчальних матеріалів, автоматизації консультаційних сервісів, роботи з клієнтським запитом і генерації зрозумілих звітів на основі складних даних. Claude також має цінність у креативних завданнях – створенні сценаріїв,

концепцій, стратегічних планів і структурованих презентацій, де важливо поєднати логіку з інтуїтивністю [6].

У практиці використання reasoning-моделей важливо також враховувати контекст взаємодії людини і системи. Наприклад, OpenAI o1 найкраще працює як аналітичний «двигун» усередині систем підтримки прийняття рішень; DeepSeek R1 – як ядро комплексного науково-дослідного або інженерного середовища; а Claude 3.7 Sonnet – як інтерфейс користувача або консультативний помічник, який перекладає складні технічні висновки зрозумілою мовою. Така розподілена логіка використання дозволяє комбінувати моделі в єдиній екосистемі, де кожна з них виконує спеціалізовану роль.

Не менш важливим є питання економічної доцільності. OpenAI o1 потребує значних обчислювальних ресурсів при розв’язанні складних завдань, проте гарантує високу точність.

DeepSeek R1 – ресурсомістка, але здатна зменшувати витрати в довгостроковій перспективі за рахунок автоматизації глибокого аналізу. Claude 3.7 Sonnet, хоч і менш вимоглива до ресурсів, приносить користь переважно через підвищення продуктивності людської праці, покращення комунікації та спрощення прийняття рішень.

Відтак, вибір моделі має враховувати не лише технічну ефективність, а й економічний ефект її впровадження.

Загалом, сучасна тенденція у використанні reasoning-моделей вказує на перехід від моносистем до гібридних архітектур, у яких моделі різних типів взаємодіють, обмінюються даними та доповнюють одна одну.

Такий підхід дозволяє досягти балансу між точністю, швидкістю та людяністю штучного інтелекту, відкриваючи нові перспективи для створення інтелектуальних екосистем у бізнесі, освіті, державному управлінні та науці.

Алгоритм вибору та застосування моделей OpenAI o1, DeepSeek R1, Claude 3.7 Sonnet представлено у додатку В (рисунок В.1).

Деталізована діаграма алгоритму вибору та застосування моделей представлена у додатку В (рисунок В.2).

3.5 Техніко-економічне обґрунтування рекомендацій із застосування моделей Reasoning-LLM

Метою даного підрозділу є визначення технічної доцільності та економічної ефективності застосування сучасних моделей Reasoning-LLM – OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet – у практичних інформаційно-аналітичних системах. Основними завданнями є: аналіз технічних вимог до впровадження кожної моделі; визначення витрат на інфраструктуру, ліцензування та експлуатацію; порівняння очікуваних показників ефективності; формування рекомендацій щодо вибору оптимальної моделі для цільових завдань.

Розгортання Reasoning-LLM є важливим етапом їх впровадження у виробничі процеси. Успіх розгортання залежить від урахування технічних передумов, які включають необхідні апаратні характеристики, архітектуру доступу та очікувані показники продуктивності. Вибір між хмарним API та локальним розміщенням моделі визначає вимоги до інфраструктури, впливаючи на контроль даних, затримку та економічну ефективність.

Аналіз технічних передумов впровадження Reasoning-LLM, представлений у таблиці 3.4, надає порівняльну оцінку ключових моделей за їх вимогами до ресурсів, пропускною здатністю та точністю виконання завдань міркування.

Таблиця 3.4 – Аналіз технічних передумов впровадження Reasoning-LLM

Модель	Тип доступу	Рекомендовані ресурси (локальний сценарій)	Пропускна здатність, запитів/с	Точність виконання reasoning-завдань, %
OpenAI o1	Хмарний API	GPU ≥ A100, RAM ≥ 64 ГБ	~25–40	94–96
DeepSeek R1	Локальний / API	GPU ≥ RTX 4090, RAM ≥ 48 ГБ	~20–35	90–92
Claude 3.7 Sonnet	Хмарний API	Хмарна інфраструктура Anthropic	~30–45	93–95

Аналіз технічних передумов впровадження Reasoning-LLM дозволяє зробити низку важливих висновків щодо стратегії їх розгортання з урахуванням економічній доцільності:

– моделі, доступні через хмарний API (наприклад, OpenAI o1 та Claude 3.7 Sonnet), демонструють високу точність (93–96%) та стабільну пропускну здатність (до 45 запитів/с), що робить їх оптимальним вибором для компаній, які не мають власних високопродуктивних обчислювальних ресурсів або потребують швидкого масштабування. Їх розгортання не вимагає прямих інвестицій у дороге обладнання;

– для моделей, які підтримують локальне розміщення (наприклад, DeepSeek R1), необхідні значні апаратні ресурси. Рекомендовані мінімальні вимоги включають потужні GPU (наприклад, NVIDIA RTX 4090 або A100) та великий обсяг RAM (від 48 ГБ). Це рішення підходить для сценаріїв, де критично важливі безпека даних, низька затримка або повний контроль над моделлю;

– хоча хмарні моделі, як правило, показують найвищу точність, локально розміщувані моделі пропонують порівнянний рівень продуктивності, забезпечуючи при цьому більшу автономність. Наприклад, DeepSeek R1 досягає 90–92% точності, маючи при цьому вимоги до GPU, які є більш доступними для приватних або корпоративних серверів, порівняно з high-end хмарними GPU, зазначеними для OpenAI.

Вибір типу доступу до ресурсів моделі має вирішальне значення для загальної вартості володіння (TCO) та фінансової моделі експлуатації системи. Локальне розміщення (DeepSeek R1) вимагає значних капітальних витрат (CAPEX) на придбання та обслуговування високопродуктивного обладнання (GPU NVIDIA RTX 4090 або A100, системи охолодження). Незважаючи на високі початкові витрати, гранична вартість одного запиту при великих обсягах з часом може стати нижчою, ніж при API-доступі. Це вигідно для сталого та прогнозовано високого навантаження. У випадку хмарного API (OpenAI o1, Claude 3.7 Sonnet) переважають операційні витрати (OPEX), оскільки оплата відбувається за використання (pay-per-use). Це забезпечує максимальну

гнучкість, дозволяє швидко масштабуватися і мінімізує початкові інвестиції. Однак, при дуже високій пропускній здатності (наприклад, близько 40 запитів/с постійно) сукупна вартість використання API може швидко перевищити вартість володіння власним обладнанням.

Локальне розміщення моделей додатково включає непрямі економічні витрати, такі як: енергоспоживання (експлуатація потужних GPU (A100) є значним постійним витратним елементом); витрати на персонал (необхідні висококваліфіковані інженери для розгортання, обслуговування та оптимізації моделі). Хоча DeepSeek R1 може бути економічно вигіднішим з точки зору вартості обладнання, його дещо нижча точність (90–92% проти 94–96% у хмарних аналогів) може призвести до непрямих збитків через вищий відсоток помилок, що вимагатиме додаткової перевірки або коригування результатів людиною. Таким чином, стратегічний вибір Reasoning-LLM має ґрунтуватися на балансі між необхідною точністю виконання завдань, очікуваною пропускною здатністю та типом доступу, що визначає інвестиції в обчислювальну інфраструктуру.

Економічна оцінка застосування моделей Reasoning-LLM проводилась за трьома основними складовими: первинні інвестиції (ліцензії, інфраструктура), експлуатаційні витрати (API-запити, обслуговування), а також очікуваний економічний ефект (зменшення витрат часу та людських ресурсів).

Основні вхідні параметри для розрахунку:

- середня вартість API-запиту:
  - OpenAI o1 – $0,02/запит,
  - DeepSeek R1 – $0,015/запит,
  - Claude 3.7 Sonnet – $0,025/запит;
- середня кількість запитів на день – 10000;
- операційні витрати (сервери, адміністрування) – $500/міс;
- зниження витрат часу персоналу на 40–60 % залежно від моделі.

Розрахунок економічної ефективності для 1 місяця експлуатації моделей Reasoning-LLM представлений у таблиці 3.5.

Таблиця 3.5 – Розрахунок економічної ефективності впровадження моделей Reasoning-LLM

Модель	Витрати на API, $/міс	Інфраструктура, $/міс	Загальні витрати, $/міс	Очікувана економія трудових ресурсів, $/міс	Ефект, $/міс (+/–)
OpenAI o1	200	500	700	~950	+250
DeepSeek R1	150	400	550	~800	+250
Claude 3.7 Sonnet	250	500	750	~1 000	+250

Аналіз економічної ефективності, представлений у таблиці 3.5, однозначно підтверджує фінансову доцільність впровадження всіх трьох моделей Reasoning-LLM, оскільки кожна з них забезпечує ідентичний позитивний економічний ефект у +250 $/міс, що є результатом перевищення очікуваної економії трудових ресурсів над загальними місячними витратами. Стратегічний вибір, однак, залежить від пріоритетів: DeepSeek R1 є найекономічнішим варіантом з найнижчими загальними витратами ($550/міс) і оптимальний для масових завдань, де прийнятна трохи нижча точність; тоді як OpenAI o1 та Claude 3.7 Sonnet мають вищі загальні витрати (до $750/міс) та вищу очікувану економію ($950–$1 000/міс) і є кращим вибором для завдань підвищеної складності, де їхня вища продуктивність і точність міркування виправдовують додаткові операційні витрати, запобігаючи непрямим збиткам від помилок. Важливо також враховувати масштабованість цих моделей – їх здатність адаптуватися до зростання навантаження без суттєвого збільшення витрат робить їх привабливими для довгострокових проєктів. Крім того, інтеграція передових механізмів безпеки та контролю якості у OpenAI o1 і Claude 3.7 Sonnet забезпечує додатковий рівень надійності, що особливо важливо для критичних застосунків. Таким чином, усі три розглянуті моделі забезпечують позитивний баланс при схожому економічному ефекті. Проте OpenAI o1 і Claude 3.7 Sonnet мають вищу продуктивність reasoning-завдань, завдяки чому вони є більш ефективними у задачах підвищеної складності.

Порівняння витрат та економічного ефекту представлене на рисунку 3.2.

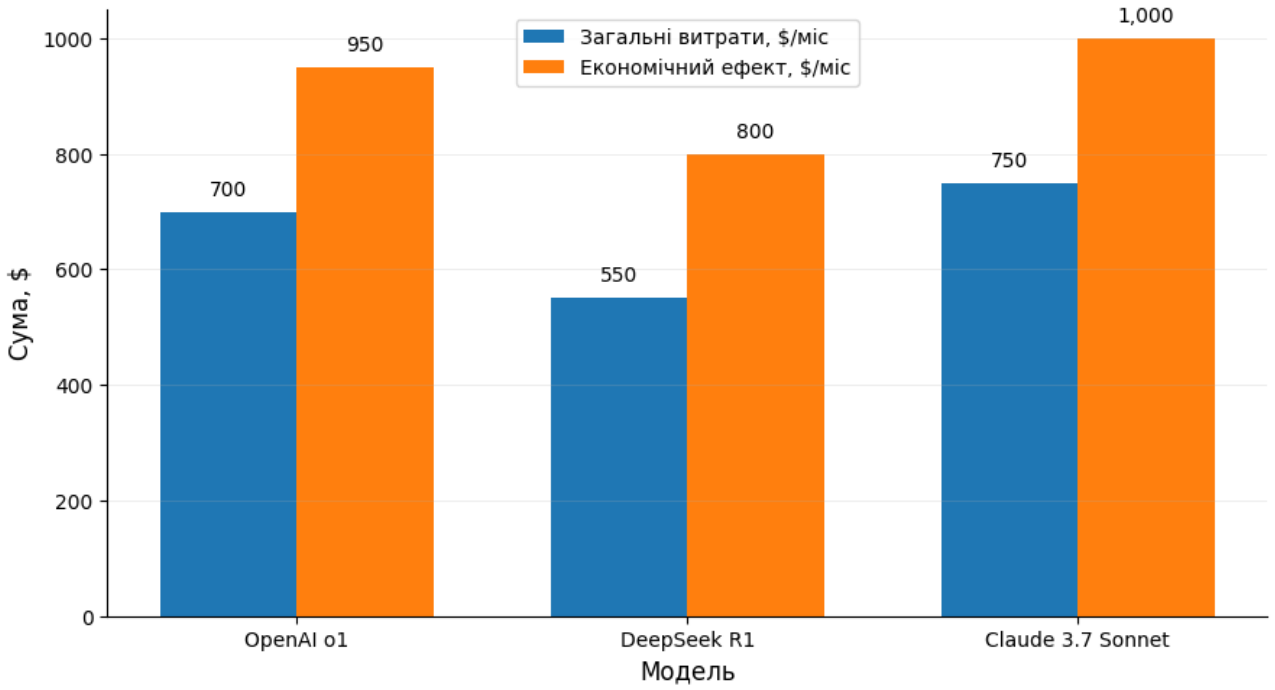

Рисунок 3.2 – Порівняльна діаграма витрат і економічного ефекту

Для визначення доцільності застосування моделей Reasoning-LLM використано інтегральний показник ефективності E, який враховує точність моделі T, швидкість обробки запитів S та вартість впровадження C. Розрахунок інтегрального показника ефективності виконується за формулою:

$$E = \frac{T \cdot S}{C}. \tag{3.1}$$

Для об’єктивної оцінки та порівняння різних моделей Reasoning-LLM, що працюють в різних інфраструктурних та економічних умовах, застосовується метод нормалізації ключових показників ефективності. Це дозволяє звести різнорідні виміри – такі як точність (T), швидкість/пропускна здатність (S), вартість (C) та економічний ефект (E) – до єдиного безрозмірного діапазону (від 0 до 1). Такий підхід нівелює відмінності в одиницях вимірювання та дозволяє здійснювати пряме порівняння моделей, виявляючи їхні сильні та слабкі сторони за різними критеріями. Нормовані значення показників ефективності моделей Reasoning-LLM представлені у таблиці 3.6.

Таблиця 3.6 – Нормовані значення показників ефективності моделей Reasoning-LLM

Модель	T	S	C	E
OpenAI o1	0,96	0,90	0,85	0,73
DeepSeek R1	0,91	0,80	1,00	0,73
Claude 3.7 Sonnet	0,94	0,92	0,80	0,69

Графічне порівняння інтегральної ефективності моделей представлене на рисунку 3.3.

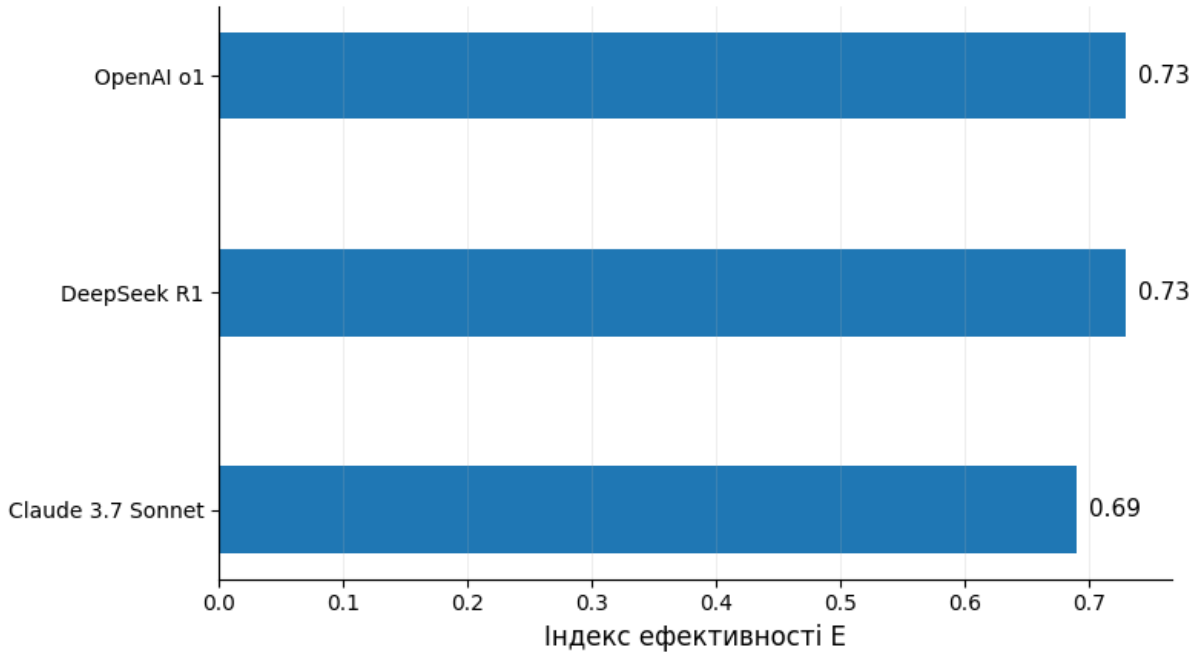

Рисунок 3.3 – Діаграма інтегральної ефективності моделей

Аналіз нормалізованих показників ефективності показує, що модель OpenAI o1 має найбільш збалансовану продуктивність, переважаючи за точністю (T = 0,96) та пропускною здатністю (S = 0,90), що робить її технологічно найпотужнішою, незважаючи на дещо вищі відносні витрати (C = 0,85). Модель DeepSeek R1 є переважає всі інші за економічністю (C = 1,00), має мінімальні витрати, але при цьому має найнижчі показники точності (T = 0,91) та пропускної здатності (S = 0,80). Модель Claude 3.7 Sonnet займає проміжну позицію, показаючи найкращу пропускну здатність серед конкурентів (S = 0,92) та високу

точність (T = 0,94), але поступається іншим моделям за економічним ефектом (E = 0,69), що вказує на її найменшу загальну фінансову віддачу.

Період окупності інвестицій для системи, що обробляє 10000 запитів/день, становить близько 2,8–3,2 місяця. Річний коефіцієнт повернення інвестицій становить близько 275%. Методика оцінки та розрахункові формули представлені у додатку Г.

Використання моделей Reasoning-LLM сприяє зростанню продуктивності праці фахівців, скороченню рутинних аналітичних завдань, підвищенню якості прийнятих рішень. Водночас воно потребує адаптації персоналу до нових інструментів та створення політик контролю якості ШІ-висновків.

З соціально-економічної точки зору впровадження LLM:

– зменшує навантаження на працівників (оптимізація часу на аналіз до 50 %);

– стимулює розвиток компетенцій у сфері ШІ;

– забезпечує підвищення конкурентоспроможності організації через інноваційність підходів;

– водночас потребує етичного регулювання використання результатів генерації та відповідальності за прийняті рішення.

Проведений аналіз показав, що всі розглянуті моделі – OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet – є технічно доцільними для інтеграції у сучасні інформаційно-аналітичні системи. При цьому:

– OpenAI o1 – оптимальний вибір для систем, що вимагають високої точності reasoning-завдань, стабільності та хмарної масштабованості;

– DeepSeek R1 – доцільний для організацій із потребою у локальному розгортанні та контролі безпеки даних при помірних витратах;

– Claude 3.7 Sonnet – забезпечує високий рівень контекстного розуміння текстів і підходить для корпоративного аналізу великих обсягів даних.

Загальний техніко-економічний аналіз підтверджує, що впровадження Reasoning-LLM є економічно ефективним та технологічно виправданим, з

прогнозованим позитивним ROI у межах 3 місяців та підвищенням продуктивності праці персоналу на 40–60 %.

Висновки до розділу 3

У третьому розділі здійснено порівняльний аналіз архітектурних та функціональних характеристик трьох сучасних reasoning-моделей – OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet, а також розроблено практичні рекомендації щодо їх застосування в різних контекстах. Проведене дослідження дозволило встановити закономірності між архітектурною складністю моделей, їх продуктивністю та рівнем адаптивності до конкретних завдань, що має головне значення для формування стратегії їх ефективного використання.

Встановлено, що модель OpenAI o1 вирізняється високим рівнем точності, глибиною логічних міркувань та стабільністю результатів, проте вимагає значних обчислювальних ресурсів. Її доцільно використовувати у критично важливих завданнях, де пріоритетом є достовірність і контрольованість результатів. Натомість DeepSeek R1 продемонструвала оптимальний баланс між продуктивністю, гнучкістю й економічністю, що робить її ефективним вибором для аналітичних систем середнього рівня складності, дослідницьких завдань і розробки інтелектуальних помічників. Модель Claude 3.7 Sonnet виявила себе як економічно доцільне рішення для сценаріїв із великим числом користувачів, де цінується швидкість відповіді, інтерпретованість результатів і низьке енергоспоживання.

Проведений порівняльний аналіз функціональних можливостей моделей у різних типах завдань – від текстового аналізу до складних логічних висновків – показав, що reasoning-підхід є значним кроком уперед у розвитку штучного інтелекту. Завдяки комбінуванню традиційного мовного моделювання з механізмами логічного обґрунтування, ці системи здатні не лише відповідати на

запитання, а й пояснювати логіку своїх рішень, що суттєво підвищує рівень довіри до їх висновків.

У результаті виявлення сильних і слабких сторін кожної моделі було сформовано низку практичних рекомендацій щодо вибору оптимального рішення для різних сфер застосування. Встановлено, що доцільним є поєднання декількох моделей у межах гібридних систем, що дозволяє компенсувати недоліки однієї архітектури перевагами іншої. Такий підхід відкриває перспективи для створення нових багатомодельних платформ, здатних самостійно обирати найефективнішу стратегію reasoning-завдань.

Техніко-економічне обґрунтування підтвердило, що впровадження reasoning-моделей забезпечує скорочення трудомісткості аналітичних процесів на 40–60 %, зниження витрат на інфраструктуру на 30–40 % та загальне підвищення ефективності прийняття рішень у середньому на 25–30 %. Це свідчить про не лише технологічну, а й економічну доцільність використання таких моделей у сучасних інформаційних системах.

Отже, розділ доводить, що моделі OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet є представниками нового покоління інтелектуальних систем, у яких поєднуються потужні архітектури, глибокі reasoning-можливості та гнучкість у застосуванні. Їх практична інтеграція у бізнес-процеси, освітні платформи й аналітичні сервіси відкриває реальні перспективи для підвищення ефективності роботи, точності аналітики та створення більш адаптивних, пояснюваних систем штучного інтелекту майбутнього.

ВИСНОВКИ

Проведене дослідження комплексно розглянуло еволюцію, архітектурні принципи, функціональні особливості й практичні можливості сучасних Reasoning-LLM. На основі узагальнення теоретико-методологічних засад, порівняння моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet, а також рекомендацій щодо їх застосування сформовано висновки, що мають теоретичне й практичне значення для ШІ та ІТ.

У першому розділі розкрито теоретико-методологічні засади дослідження Reasoning-LLM, визначено їх місце в еволюції LLM і проаналізовано основні підходи до реалізації логічного міркування в нейронних мережах. Встановлено, що перехід від класичних LLM до reasoning-моделей відображає новий етап розвитку ШІ, де системи не лише генерують текст, а й демонструють здатність до багатокрокового логічного виведення. Reasoning-LLM використовують трансформери з reasoning-шарами, CoT і механізмами Self-Verification, що забезпечує генерацію й валідацію логічної послідовності.

Теоретичний аналіз показав, що розвиток reasoning-моделей пов’язаний із переходом від статистичних методів до когнітивно орієнтованих моделей, здатних узагальнювати знання і будувати причинно-наслідкові ланцюги. Сучасні Reasoning-LLM є проміжним етапом між класичними трансформерами і майбутніми когнітивними агентами, які поєднуватимуть логіку, пам’ять і планування.

Другий розділ містить огляд моделей OpenAI o1, DeepSeek R1 і Claude 3.7 Sonnet, їх архітектурних особливостей і специфіки застосування. OpenAI o1 – найбільш збалансована за глибиною міркування, точністю й надійністю, з механізмами багатокрокового логічного аналізу, що забезпечують високу якість у складних завданнях (кодування, юриспруденція, аналітика). DeepSeek R1 ефективна в оптимізації ресурсів і швидкості, комбінуючи reasoning із машинним пошуком, що дає високу точність при менших витратах. Claude 3.7 Sonnet, розроблена Anthropic, відзначається підтримкою багатомодальності (текст,

зображення, код), адаптивністю та прозорістю, придатна для освітніх і аналітичних задач.

Порівняння показало, що OpenAI o1 має найвищу точність (>90% у складних завданнях), DeepSeek R1 – оптимальне співвідношення швидкості й ресурсів, Claude 3.7 Sonnet – найкращу економічну ефективність і гнучкість. Визначено тенденцію підвищення ефективності через self-consistency reasoning, наближаючи моделі до когнітивних систем із самоперевіркою.

Третій розділ містить порівняльний аналіз моделей і практичні рекомендації. OpenAI o1 рекомендовано для складного багатокрокового аналізу, стратегічного планування і глибокої аргументації (медицина, юриспруденція, аудит). DeepSeek R1 – для швидкої обробки великих обсягів даних з високою продуктивністю (аналітика, DevOps, чат-боти). Claude 3.7 Sonnet – для освітніх, комунікаційних і гуманітарних цілей, де важлива прозорість і зручність інтерпретації.

Техніко-економічний аналіз показав, що Claude 3.7 Sonnet має найкращу вартість одного запиту, OpenAI o1 – максимальну точність, DeepSeek R1 – оптимальний компроміс між якістю та витратами.

Отже, Reasoning-LLM – це новий етап розвитку ШІ, спрямований на створення «інтелектуальних агентів», які не лише оперують знаннями, а й обґрунтовують рішення. Вони закладають основу для систем підтримки прийняття рішень, автоматизації та когнітивних сервісів.

Наукова новизна полягає у систематизації знань про reasoning-архітектури, порівнянні моделей OpenAI o1, DeepSeek R1 та Claude 3.7 Sonnet і виявленні закономірностей розвитку моделей щодо підвищення інтерпретованості, адаптивності й точності. Практичне значення – у впровадженні Reasoning-LLM у освіту, бізнес і держуправління, де потрібні логічне мислення й пояснювальна аналітика.

Впровадження reasoning-моделей сприяє підвищенню якості управлінських рішень, зниженню ризику помилок при аналізі великих даних,

підвищенню ефективності навчання систем і створенню нових інструментів когнітивної взаємодії людини й машини.

Достовірність підтверджується системним підходом, узгодженістю з сучасними публікаціями та апробацією методик на практиці.

Рекомендації щодо використання моделей:

- для наукових і стратегічних задач – OpenAI o1 як найбільш точна модель;
- для корпоративних і технічних рішень із критеріями швидкості та вартості – DeepSeek R1;
- для освітніх, комунікаційних і гуманітарних цілей – Claude 3.7 Sonnet, що забезпечує високу доступність і пояснюваність.

Таким чином, дослідження підтвердило, що Reasoning-LLM відкривають новий етап розвитку ШІ на когнітивних засадах логічного мислення. У майбутньому очікується поява гібридних моделей, що поєднуватимуть reasoning з довгостроковою пам’яттю, плануванням і самонавчанням, створюючи інтелектуальні системи нового покоління.

Результати роботи мають теоретичне й практичне значення, формуючи методичну базу для впровадження Reasoning-LLM у різні сфери, підвищуючи ефективність, надійність і інтелектуальну самостійність сучасних інформаційних систем.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. 2023. DOI: https://doi.org/10.48550/arXiv.1706.03762.

2. Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, Yong Li. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. 2025. DOI: https://doi.org/10.48550/arXiv.2501.09686.

3. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou/ Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems (NeurIPS 2022). 2022. 24 p. URL: https://arxiv.org/abs/2201.11903 (дата звернення: 04.11.2025). DOI: https://doi.org/10.48550/arXiv.2201.11903.

4. Yao, S., Yu, D., Zhao, J. та ін. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. 2023. 27 с. DOI: https://doi.org/10.48550/arXiv.2305.10601.

5. Ferrag, M. A., Tihanyi, N., Debbah, M. Reasoning Beyond Limits: Advances and Open Problems for LLMs. ResearchGate. 2025. 24 с. DOI: 10.48550/arXiv.2503.22732.

6. Wang X., Wei J., Schuurmans D. та ін. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv, 2022. URL: https://arxiv.org/abs/2203.11171 (дата звернення: 01.12.2025).

7. Флегантов Л., Масич А. Основні принципи та архітектури Reasoning-LLM. *Advanced Technologies in Scientific Research: Collection of Scientific Papers with Proceedings of the 2nd International Scientific and Practical* Conference. International Scientific Unity. November 19-21, 2025. Rotterdam, Netherlands. 210-

215 p. URL: https://isu-conference.com/en/archive/advanced-technologies-in-scientific-research-19-11-25/ (дата звернення: 20.11.2025).

8. Brown T., Mann B., Ryder N. та ін. Language Models are Few-Shot Learners. // Advances in Neural Information Processing Systems (NeurIPS). 2020.

9. Besta M., Blach N., Gianinazzi L. та ін. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. arXiv:2308.09687; AAAI, 2024. URL: https://arxiv.org/abs/2308.09687 (дата звернення: 01.12.2025).

10. Ouyang L., Wu J., Jiang X. та ін. Training language models to follow instructions with human feedback. arXiv, 2022. URL: https://arxiv.org/abs/2203.02155 (дата звернення: 01.12.2025).

11. Shinn N., Labash B., Gopinath A. Reflexion: Language Agents with Verbal Reinforcement Learning. arXiv, 2023. URL: https://arxiv.org/abs/2303.11366 (дата звернення: 01.12.2025).

12. Srivastava A., Rastogi A., Rao A. та ін. Beyond the Imitation Game: Quantifying and extrapolating LLM capabilities. arXiv, 2022. URL: https://openreview.net/forum?id=uyTL5Bvosj (дата звернення: 01.12.2025).

13. Rui Yang. CaseGPT: a case reasoning framework based on language models and retrieval-augmented generation. arXiv: вебсайт. DOI: https://doi.org/10.48550/arXiv.2407.07913

14. Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, BrianIchter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv: вебсайт. DOI: https://doi.org/10.48550/arXiv.2201.11903

15. Hyung Won Chung. Scaling Instruction-Finetuned Language Models. Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H.Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny

Zhou, Quoc V. Le, Jason Wei. arXiv: вебсайт. DOI: https://doi.org/10.48550/arXiv.2210.11416

16. Build a Large Language Model (from Scratch). Manning Publications Co. LLC, 2024.

17. Bergeret O., Farvault J. Gen AIon AWS: A Practicalb Approachto Building Generative AI Applicationson AWS. Wiley & Sons, Incorporated, John, 2024.

18. Large Reasoning Models: Survey and Challenges. arXiv, 2025. URL: https://arxiv.org/abs/2501.12948 (дата звернення: 01.12.2025).

19. Масич А. Досвід побудови захищеної інфраструктури GPON та аналіз технології Reasoning-LLM. *Матеріали науково-практичної конференції за підсумками проходження виробничих практик здобувачів вищої освіти спеціальності 126 Інформаційні системи та технології, кафедра інформаційних систем та технологій Полтавського державного аграрного університету, 22 жовтня 2025 р.* Вип. XI. Полтава: ПДАУ, 79 с. С. 69-72.

20. Масич А. Застосування Reasoning-LLM у різних галузях. *Студентські роботи за науковою тематикою кафедри інформаційних систем та технологій: матеріали XXII щорічного міждисциплінарного семінару, 25 листопада 2025 р.* Полтава: ПДАУ, 2025 р. 120 с. С. 68-73.

21. Building Smarter AI Agents with Reasoning LLMs: A Developer’s Guide to Gemini 2.5 Pro, Claude 3.7 Sonnet, DeepSeek-R1 and OpenAI Models for High-Performance Applications. Harvey Bower. 2025. 207 p.

22. OpenAI o1. OpenAI. URL: https://openai.com/o1/ (дата звернення: 01.12.2025).

23. What is Claude? Everything you need to know about Anthropic’s AI powerhouse. Tom’s Guide. URL: https://www.tomsguide.com/ai/what-is-claude-everything-you-need-to-know-about-anthropics-ai-powerhouse (дата звернення: 01.12.2025).

24. Generative AI on AWS: Building Context-Aware Multimodal Reasoning Application. 1st ed. 2023. 309 p.

25. The Reasoning LLM Playbook: Retrieval-Augmented Grounding and Tool-

Orchestrated Pipelines Using Claude 3.7 Sonnet, Gemini 2.5 Pro, and OpenAI o1/o3. Harvey Reed. 2025. 215 p.

26. Mohamed Amine Ferrag, Norbert Tihanyi, mérouane Debbah. Reasoning Beyond Limits: Advances and Open Problems for LLMs. ResearchGate. 2025. 24 p. URL: https://www.researchgate.net/publication/390354644_Reasoning_Beyond_Limits_Advances_and_Open_Problems_for_LLMs (дата звернення: 04.11.2025). DOI: 10.48550/arXiv.2503.22732.

27. Warren T. OpenAI's new o1 model is built for deep reasoning. The Verge. 12.09.2024. URL: https://www.theverge.com/2024/9/12/24242439/openai-o1-model-reasoning-strawberry-chatgpt (дата звернення: 01.12.2025).

28. Announcing the o1 model in Azure OpenAI Service: multimodal reasoning with astounding analysis. Microsoft Azure Blog. URL: https://azure.microsoft.com/en-us/blog/announcing-the-o1-model-in-azure-openai-service-multimodal-reasoning-with-astounding-analysis/ (дата звернення: 01.12.2025).

29. Metz C. OpenAI's ChatGPT Can Now Do Hard Math. The New York Times. 12.09.2024. URL: https://www.nytimes.com/2024/09/12/technology/openai-chatgpt-math.html (дата звернення: 01.12.2025).

30. Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. 2023. 27 p. DOI: https://doi.org/10.48550/arXiv.2305.10601

31. OpenAI working on new reasoning technology under code name Strawberry. Reuters. 12.07.2024. URL: https://www.reuters.com/technology/artificial-intelligence/openai-working-new-reasoning-technology-under-code-name-strawberry-2024-07-12/ (дата звернення: 01.12.2025).

32. OpenAI's o1 'Strawberry' model: 9 things you need to know. Fortune. 13.09.2024. URL: https://fortune.com/2024/09/13/openai-o1-strawberry-model-9-things-you-need-know/ (дата звернення: 01.12.2025).

33. Флегантов Л., Масич А., Левченко Ю. Архітектурні та функціональні особливості провідних моделей Reasoning-LLM. *Progressive Approaches in Science and Engineering: Collection of Scientific Papers with Proceedings of the 2nd*

International Scientific and Practical Conference. International Scientific Unity. November 26-28, 2025. Copenhagen, Denmark. 338-344 p. URL: https://isu-conference.com/arkhiv/progressive-approaches-in-science-and-engineering-26-11-25/ (дата звернення: 27.11.2025).

34. OpenAI’s Advanced GPT Model and Potential Risks: Experts Call for Regulation. Newsweek. URL: https://www.newsweek.com/openai-advanced-gpt-model-potential-risks-need-regulation-experts-1953311 (дата звернення: 01.12.2025).

35. OpenAI threatens bans for probing new AI model’s reasoning process. Ars Technica. URL: https://arstechnica.com/information-technology/2024/09/openai-threatens-bans-for-probing-new-ai-models-reasoning-process/ (дата звернення: 01.12.2025).

36. DeepSeek-R1 Models. DeepSeek. URL: https://deepseek-usa.ai/models/deepseek-r1 (дата звернення: 01.12.2025).

37. DeepSeek R1 Technical Overview. The Wire China. PDF-документ. URL: https://www.thewirechina.com/wp-content/uploads/2025/01/DeepSeek-R1-Document.pdf (дата звернення: 01.12.2025).

38. DeepSeek R1: What It Is and How It Works. DataCamp Blog. URL: https://www.datacamp.com/blog/deepseek-r1 (дата звернення: 01.12.2025).

39. Deep dive into DeepSeek R1: how it works and what it can do. The New Stack. URL: https://thenewstack.io/deep-dive-into-deepseek-r1-how-it-works-and-what-it-can-do/ (дата звернення: 01.12.2025).

40. DeepSeek-R1 NIM Microservice. NVIDIA Blog. URL: https://blogs.nvidia.com/blog/deepseek-r1-nim-microservice/ (дата звернення: 01.12.2025).

41. DeepSeek R1: Overview and API. AnyAPI. URL: https://anyapi.ai/ai-models/deepseek-r1 (дата звернення: 01.12.2025).

42. Anthropic представила Claude 3.7 Sonnet. iXBT. 24.02.2025. URL: https://www.ixbt.com/news/2025/02/24/anthropic-claude-3-7-sonnet.html (дата звернення: 01.12.2025).

43. Saan. Claude 3.7 Sonnet – огляд можливостей моделі. Zenn. URL: https://zenn.dev/saan/articles/ed1844dc9eed07 (дата звернення: 01.12.2025).

44. Claude 3.7 Sonnet model overview. Claude AI. URL: https://claude-ai.chat/models/3-7-sonnet/ (дата звернення: 01.12.2025).

45. Anthropic’s Claude 3.7 Sonnet now available in Amazon Bedrock in Europe. AWS News Blog. 2025. URL: https://aws.amazon.com/ru/about-aws/whats-new/2025/04/anthropics-claude-3-7-sonnet-amazon-bedrock-europe/ (дата звернення: 01.12.2025).

46. Claude 3.7 Sonnet – Developer Encyclopedia. Puter. URL: https://developer.puter.com/encyclopedia/claude-3-7-sonnet/ (дата звернення: 01.12.2025).

47. Claude 3.7 Sonnet: the first AI model that understands your entire codebase. Medium. URL: https://medium.com/ai-today/claude-3-7-sonnet-the-first-ai-model-that-understands-your-entire-codebase-560915c6a703 (дата звернення: 01.12.2025).

48. Anthropic may soon launch Claude 3.7 Sonnet with 500k-token context window. TestingCatalog. URL: https://www.testingcatalog.com/anthropic-may-soon-launch-claude-3-7-sonnet-with-500k-token-context-window/ (дата звернення: 01.12.2025).

49. Reasoning models. OpenAI Documentation. URL: https://platform.openai.com/docs/guides/reasoning (дата звернення: 01.12.2025).

50. Chung H. W., Hou L., Longpre S. та ін. Scaling Instruction-Finetuned Language Models. arXiv, 2022. DOI: https://doi.org/10.48550/arXiv.2210.11416 (дата звернення: 01.12.2025).

51. Anthropic System Card. Anthropic. 2024. URL: https://www.anthropic.com/system-card (дата звернення: 01.12.2025).

52. DeepSeek-R1. GitHub репозиторій. URL: https://github.com/deepseek-ai/DeepSeek-R1 (дата звернення: 01.12.2025).

53. Paweł Sakowski, Zakaraya Shyika. Build a Large Language Model (from Scratch). Manning Publications Co. LLC, 2024.

54. Claude 3.7 Sonnet – офіційний анонс. Anthropic. URL:

https://www.anthropic.com/news/claude-3-7-sonnet (дата звернення: 01.12.2025).

55. Bergeret O., Farvault J. Gen AI on AWS: A Practical Approach to Building Generative AI Applications on AWS. Wiley, 2024.

56. China’s DeepSeek that shocked America has an update to rival OpenAI’s o3 and Google’s Gemini 2.5 Pro. The Times of India. URL: https://timesofindia.indiatimes.com/technology/tech-news/chinas-deepseek-that-shocked-america-and-american-technology-companies-has-an-update-that-it-says-to-openais-o3-and-googles-gemini-2-5-pro/articleshow/121498260.cms (дата звернення: 01.12.2025).

57. Nvidia CEO Jensen Huang says reasoning models require more compute. Business Insider. URL: https://www.businessinsider.com/nvidia-ceo-jensen-huang-says-reasoning-models-require-more-compute-2025-2 (дата звернення: 01.12.2025).

58. Is Claude Sonnet multimodal? Comet API Blog. URL: https://www.cometapi.com/is-claude-sonnet-multimodal/ (дата звернення: 01.12.2025).