Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«РОЗРОБКА ЧАТ-БОТУ ДЛЯ РОЗВ'ЯЗАННЯ МАТЕМАТИЧНИХ ЗАДАЧ НА ОСНОВІ LLEMMA / LLEMMA-BASED ASSISTANT CHAT BOT FOR SOLVING MATHEMATICAL PROBLEMS»**

Виконала: студентка 4-го року навчання,
Спеціальності
113 Прикладна математика
Сапко Марія Сергіївна
Керівник ___Кашпіровський О.І.,__
доцент,
кандидат фізико-математичних наук
Рецензент _____
Кваліфікаційна робота захищена
з оцінкою _____
Секретар ЕК _____
«___» _____20___р.

Київ – 2024

Міністерство освіти і науки України

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Зав.кафедри математики,

Доцент, к.н

_____Чорней Р.К.___

“__”_____2024 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на кваліфікаційну роботу

студентці Сапко Марії Сергіївні факультету інформатики 4 курсу

ТЕМА: «Розробка чат-боту для розв'язання математичних задач на основі llemma / llemma-based assistant chat bot for solving mathematical problems»

Зміст ТЧ до кваліфікаційної роботи:

Індивідуальне завдання

Анотація

Вступ

Розділ 1. Дослідження концепції чат-бота та моделі LLEMMA

Розділ 2. Аналіз моделі llemma_7b

Розділ 3. Розробка чат-бота

Висновок

Список використаних джерел

Дата видачі “__”_____ 2024 р. Керівник _____

(підпис)

Завдання отримав _____

(підпис)

Тема: «Розробка чат-боту для розв'язання математичних задач на основі llemma / llemma-based assistant chat bot for solving mathematical problems»

Календарний план виконання роботи:

№ п/п	Назва етапу дипломної роботи	Термін виконання етапу	Примітка
1.	Отримання завдання на курсову роботу	31.10.2023	
2.	Огляд теми та розробка плану виконання роботи	12.01.2024	
3.	Огляд технічної літератури за темою роботи	січень-лютий 2024	
4.	Написання теоретичної частини роботи	Лютий-березень 2024	
5.	Реалізація практичної частини	Березень-квітень 2024	
6.	Написання текстової частини роботи	Квітень-травень 2024	
7.	Захист дипломної роботи	3-4 червня 2024	

Студентка: Сапко М.С

Керівник: Кашпіровський О.І.

“31” жовтня 2023

Зміст

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ ТА СКОРОЧЕНЬ

ШІ – Штучний інтелект

NLP – Natural Language Processing

ML – Machine Learning

LLM – Large Language Model

HTML – HyperText Markup Language

CSS – Cascading Style Sheets

NLU – Natural Language Understanding

NLG – Natural Language Generation

FAQ – Frequently asked questions

RLHF – Reinforcement Learning from Human Feedback

PPO – Proximal Policy Optimization

GPT – Generative Pre-trained Transformer

LaMDA – Language Model for Dialogue Applications

PaLM – Pathways Language Model

IT – Information Technology

MMLU – Massive Multitask Language Understanding

Анотація

Курсова робота має на меті проведення якісного, кількісного та порівняльного аналізів моделі llemma_7b, створеної для вирішення математичних задач, та розробку чат-бота на її основі. Було створено набір завдань різної складності, згрупованих за розділами математики, для проведення обширного комплексного тестування. Для проведення порівняльного аналізу було обрано найбільш популярні чат-боти станом на березень 2024 року. Основну частину застосунку було створено на мові Python із використанням фреймворка Flask , front-end частину – мовами HTML, CSS, JavaScript. Було реалізовано основні функції прийому запиту користувача, його відправка моделі, обробка відповіді моделі та демонстрація результату користувачу.

Ключові слова: чат-бот, модель, llemma_7b, аналіз, ефективність

ВСТУП

«Штучний інтелект», «машинне навчання», «автоматизований онлайн-помічник» - слова які сучасна пересічна людина чує кожного дня. Новітні технології задіяні у багатьох сферах нашого життя і ми радо користуємось ними, адже вони спрощують різноманітні процеси, відкривають нові можливості та економлять один із найважливіших людських ресурсів – час. Інтелектуальні помічники у вигляді чат-ботів стали однією із таких інновацій. Вони набули поширення у різних галузях, включаючи освіту, оскільки із популяризацією самостійного навчання стрімко зріс і попит на «віртуальних персональних вчителів». Створення асистента на основі моделі LLEMMA є перспективним варіантом вирішення проблеми надання додаткової допомоги у розв'язанні математичних задач.

Математика має широку сферу застосування в різноманітних галузях людської діяльності, що зумовлює високий попит на фахівців цього профілю. Охочі опанувати науку нерідко стикаються із труднощами у розумінні та засвоєнні навчального матеріалу і відповідно у вирішенні певних задач. Використовуючи можливості штучного інтелекту і обробки природної мови заявлений чат-бот має на меті покращити процес навчання зацікавлених. Спеціалізуючись на вирішенні задач різних дисциплін математики, Модель Llemma надає можливість створити асистента, здатного розуміти запит користувача і надати відповідь достатньо простою, подібною до звичайної людської, мовою. Швидкий зворотній зв'язок, детальні рішення, персоналізовані вказівки та покрокові пояснення забезпечать необхідною інформацією користувача, що підвищить його обізнаність у відповідних темах і призведе до покращення навичок вирішення певних завдань. Функціонал запропонованого чат-бота передбачає як надання рішень до завдань розрахованих на базовий рівень знань, так і забезпечення допомоги для успішного розв'язання задач підвищеної складності. Наукове ж значення проекту полягає у здійснені внеску в розвиток

технологій ШІ, NLP та ML, розглядаючи їх застосування в галузі математичної освіти.

Новостворена мовна модель Llemma постає об'єктом дослідження у даній роботі. На поточному етапі розробки зазначена модель не набула широкого розповсюдження та значної популярності, у порівнянні з іншими всесвітньовідомими LLM, проте її потенціал та перспективність варті уваги розробників. Її вдосконалення та імплементація можуть стати вагомим кроком вперед у створенні допоміжних технологій у галузі математики. Відповідно предмет дослідження – процес розробки інтелектуального чат-бота на основі вищезазначеної моделі.

За основну мету даної роботи було поставлено дослідження ефективності моделі та її компетентність у виконанні згаданих вище задач, виявлення особливостей використання та створення навчального помічника на її основі. Завдяки здатності імітувати живе мовлення, чат-бот став оптимальним варіантом для втілення асистента. Для досягнення цієї мети були визначені наступні завдання:

- Визначити поняття чат-боту, його особливості та процес розробки, дослідивши відповідну літературу та наявні приклади.
- Детально розглянути новітню модель Llemma. Зібрати та попередньо підготувати дані для тестування, провести аналіз, виявити унікальні властивості притаманні моделі, щоб визначити перспективи її використання та можливості застосування.
- Спроектувати раціональну структуру програми, визначити оптимальне середовище розробки та додаткові інструменти відповідно до поставлених цілей.
- Розробити інтуїтивно зрозумілий та практичний інтерфейс чат-бота для зручного користування.

- Реалізувати основні функції помічника, такі як прийом та обробка вхідних даних користувача, формування запиту для моделі та його надсилання, форматування отриманого результату та відправка відповіді користувачу.
- Провести широке комплексне тестування та оцінити коректність результатів і повноту відповідей, ефективність роботи чат-бота, зручність використання і визначити перспективи подальшого вдосконалення програми.

Для дослідження та подальшої роботи було обрано модель Llemma_7b. З огляду на великий обсяг зазначеної моделі та необхідність опрацювання великої кількості експериментальних запитів, процес тестування моделі було здійснено в Google Collaboratory з використанням локального середовища потужного комп'ютера. Google Colab надає зручне інтерактивне середовище для розробки та налагодження коду, що полегшує процес тестування, допомагаючи організувати простір та дозволяючи коригувати та перезапускати тільки певні частини коду. В той час як обране середовище забезпечило можливість встановити модель та доволі швидко обробляти запити до неї. Основою для тестування стали завдання із навчальних посібників та згенеровані ШІ вхідні дані, додатково адаптовані для проведення більш обширного та різнопланового тестування. Аналіз, обробку та інтерпретацію результатів, а також оцінку ефективності та точності моделі було проведено без застосування додаткових інструментів.

Оптимальною мовою для розробки основного функціоналу чат-бота було визначено Python. Для створення користувацького інтерфейсу була здійснена інтеграція Python з такими мовами як JavaScript, HTML та CSS.

РОЗДІЛ 1: Дослідження концепції чат-бота та моделі LLEMMA

1.1 Загальна інформація про чат-бот

1.1.1 Означення терміна, основні аспекти, переваги та недоліки

Чат-бот – це комп'ютерна програма створена для того, щоб імітувати розмову з користувачем способом текстової або голосової взаємодії. Вона обробляє вхідний запит, наданий людською мовою, та генерує свою відповідь звертаючись до бази даних попередньо сформованих усталених відповідей або застосовуючи технології глибокого машинного навчання для надання більш різноманітної та вичерпної інформації.

Первинні чат-боти використовувались як один із способів реалізації концепції FAQ. Вони спиралися на обмежений набір поширених запитань та заздалегідь підготовлених відповідей, вимагаючи від користувачів обирати ключові слова і фрази, щоб знайти відповідну інформацію. Такі базові програми не були здатні обробляти непередбачені розробниками запити. Сучасні ж чат-боти використовують передові технології штучного інтелекту, такі як розуміння природної мови (NLU), її обробка (NLP) та генерація (NLG), машинне навчання, щоб вести більш натуральні та ситуативні діалоги. Згідно з прогнозами наступне покоління чат-ботів характеризуватиметься ще більш розширеною функціональністю, а саме здатністю виявляти стиль спілкування користувача та емоційний стан. На основі цих факторів, програма адаптуватиме свою лексику, манеру викладу інформації та тон спілкування на відповідний формат, щоб зробити процес комунікації більш персоналізованим та комфортним. Чат-боти, які використовують генеративний ШІ, можуть не тільки підтримувати розмову та передавати інформацію словами, а й доповнювати свою відповідь самостворенними зображеннями та аудіодоріжками, перекладати, узагальнювати, робити прогнози у відповідь на запит користувача.

З появою даної технології зникла проблема обмеженого часом доступу до інформації та послуг. Раніше отримати консультацію експерта можна було

тільки в робочі часи організації за наявності доступного персоналу. Наразі чат-боти миттєво реагують на повідомлення користувачів, адже вони активні 24/7 і можуть одночасно обробляти велику кількість запитів. Надаючи вичерпні відповіді та щохвилинно вдосконалюючи їх якість, віртуальні помічники звільнили працівників від виконання однотипних часто повторюваних задач. Будь-які технології можуть як стати корисним інструментом, так і нести в собі загрозу або призводити до негативних наслідків, якщо допустити помилки в їх розробці чи використовувати неналежним чином, чат-боти не стали виключенням. Моделі, які використовують ШІ, можуть нести ризики для безпеки, зокрема порушення конфіденційності або правил інтелектуальної власності, допущення витоку даних. Також неточні або нерелевантні відповіді можуть бути спричинені потраплянням недоречних даних у навчальні датасети або недостатньою освіченістю моделі. До того ж наразі чат-боти добре справляються з повторюваними завданнями легкої та середньої складності, проте не дають повну розгорнуту правильну відповідь на нестандартні, технічні та деталізовані запити, які вимагають проведення більш глибокого дослідження та виконання багатоетапного алгоритму дій. Більш ефективними в опрацюванні вузьконаправлених запитів можна вважати спеціалізовані моделі, навчені для надання допомоги лише в певній предметній області. Конкретними прикладами таких моделей можуть бути ChatSpot (допомагає у вирішенні питань продажів, маркетингу та аналітики) та Amazon CodeWhisperer (профілюється на написанні, редагуванні та оптимізації програмного коду).

Чат-боти набули широкого поширення у сферах інтернет-торгівлі, фінансів, галузі туризму, охорони здоров'я та індустрії медіа. Інтелектуальні помічники допомагають у виконанні найрізноманітніших завдань, серед найбільш поширених функцій можна виокремити:

- Оформлення замовлень, надсилання нагадувань про очікувану подію, відслідковування пакунків та супровід на кожному кроці здійснення покупки онлайн

- Здійснення грошових переказів або оплата рахунків, прогнозування майбутніх тенденцій у динаміці різноманітних фінансових активів, надання рекомендацій щодо найбільш прибуткового плану інвестування
- Організація подорожей на основі вподобань користувачів: починаючи з порад щодо варіантів проживання, закінчуючи створенням оптимального детального поденного маршрутного плану, який охопить бажані для перегляду місця.
- Встановлення діагнозу на основі симптомів пацієнта, надання повної інформації про медичну систему і порад щодо найбільш підходящого місця лікування, допомога у лікуванні психічних захворювань
- Моніторинг та швидка публікація новин, підбивання підсумків на основі певної інформації та аналіз вподобань аудиторії.

Чат-боти швидко набули поширення, а із інтеграцією технології ШІ стали невід'ємною частиною повсякденного життя сучасної людини.

1.1.2 Представники обрані для подальшого дослідження

Метою даного дослідження було оцінити ефективність та компетентність моделі, а також визначити необхідність створення помічника на її основі, шляхом порівняння із уже наявними найбільш поширеними передовими чат-ботами. Обрані програми не спеціалізуються на вирішенні математичних завдань, проте дані з даної дисципліни входили у стек матеріалів для навчання моделей. Нижче наведено перелік описаних чат-ботів та їх характерні особливості, які варто враховувати під час аналізу:

1) ChatGPT

   Це чат-бот із ШІ, розроблений OpenAI і виставлений на широкий загал 30 листопада 2022 року. За перші два тижні після запуску він зібрав понад 1 мільйон користувачів, а за наступні 2 місяці їх кількість зросла в 100 разів.

   LLM ChatGPT була навчена за допомогою методу навчання з підкріпленням на основі зворотного зв'язку від справжніх людей (RLHF). Увесь процес розробки можна умовно поділити на 3 кроки:

   1. Навчання початкової моделі: справжні люди надавали приклади очікуваних відповідей на певні запити, на яких навчалася модель
   2. Тренування моделі винагороди (reward model): працівники впорядковували відповіді моделі від найкращої до найгіршої, на чому і навчалася дана модель
   3. Навчання за алгоритмом політики проксимальної оптимізації (PPO): певна політика генерує відповідь, модель винагороди визначає винагороду, ця винагорода використовується щоб оновити політику. Цей цикл виконується кілька разів, щоб модель стала більш успішною у генеруванні послідовних відповідей.

Наразі існує дві версії чат-бота: GPT- 3.5 та GPT Plus (на основі GPT-4). У подальшому дослідженні використано першу.

2) Claude

Це сімейство великих мовних моделей(LLM) і ШІ-помічників, розроблених компанією Anthropic. В березні 2023 року була випущена перша версія Claude 1.3, у листопаді того ж року була показана нова версія Claude 2.1, що була натренована на в десятки разів більшій кількості даних, у порівнянні з попередньою моделлю. Основним завданням версії Claude 3, яка була обрана для тестування, заявлено допомогу у виконанні різноманітних завдань, таких як письмо, аналіз, кодування, математика та творчі проекти.

Інновацією Anthropic стало розроблення власного підходу «Конституційний штучний інтелект» (Constitutional AI) і його впровадження у процес навчання розроблюваних моделей. Цей підхід за концепцією дуже схожий на RLHF, проте для навчання використовує відгуки ШІ, а не людські. Використовуючи генеративні попередньо навчені трансформатори (GPT), у поєднанні з Конституційним штучним інтелектом та RLHF, розробники організації з часом підвищили прозорість (відмова від виконання шкідливих запитів) моделі та зменшили її залежність від людського нагляду. Від інших мовних моделей Claude відрізняє встановлення акценту на безпеці та здатності генерувати відповіді згідно з етичними нормами.

3) Gemini

Це чат-бот розроблений компанією Google шляхом об’єднання моделей Bard та Duet AI у лютому 2024 року. Тож спочатку він базувався на сімействі мовних моделей LaMDA, а пізніше на PaLM. З самого початку він створювався мультимодельним, тобто він може оперувати та працювати з різними типами інформації включаючи текст, код, аудіо,

зображення та відео. Існує чотири версії чат-бота: Gemini, Gemini Ultra, Gemini Pro, Gemini Nano. Основною стратегією навчання компанія визначила тренування менших моделей на більшій кількості токенів, щоб покращити продуктивність. Для навчання версії Gemini Ultra було використано 11 мільярдів токенів, для Gemini Pro – 5.5. Вихідні дані моделей під час навчання проходили перевірку у три етапи: оцінка за допомогою набору заздалегідь визначених правил та інструкцій, перевірка класифікаторами, навченими виявляти та видаляти шкідливий, неетичний або недоречний контент та додатковий крок фільтрації з орієнтацією на безпеку даних, на якому відбувається видалення шкідливого контенту.

У подальшому дослідженні фігуруватиме стандартна версія Gemini.

4) Perplexity AI

Це пошуковий чат-бот розроблений командою з 4 IT-спеціалістів та випущений у серпні 2022 року. Програма реалізована на основі таких відомих моделей, як GPT-3.5 у безкоштовній версії та GPT-4, Claude 3, Mistral Large, Llama 3 та Experimental Perplexity Model у Perplexity Pro. Головною особливістю чат-бота вважається його система пошуку, яка здійснює поглиблений інформаційний скринінг на основі запиту користувача. Розробниками заявлено, що Perplexity AI індексує Інтернет щодня, тож чат-бот доволі компетентний у наданні інформації про актуальні питання сьогодення. Разом із відповіддю помічник надає список ресурсів, які були знайдені та використані для формування вихідного повідомлення, що дозволяє глибше зануритися у дослідження, а не цілковито довіряти ШІ.

Для подальшого аналізу було обрано звичайну версію Perplexity AI.

1.2 Модель LLEMMA

1.2.1 Основні характеристики моделі

LLEMMA – це LLM, що спеціалізується на вирішенні математичних задач. Вона була розроблена некомерційною групою дослідників у галузі ШІ – EleutherAI. Об'єднання було створене у липні 2020 року на платформі Discord з метою створити реплікацію GPT-3, а вже на початку 2023 року воно було офіційно зареєстроване як EleutherAI Foundation, некомерційний дослідницький інститут. Наразі здобутки організації налічують 26 різноманітних проектів з яких 13 – розробка моделей, інші – формування датасетів та створення бібліотек. Основна особливість EleutherAI – це те, що вони надають відкритий доступ до коду кожного з проектів.

Модель LLEMMA була створена на основі моделі Code Llama шляхом проведення додаткового навчання на спеціально створеному датасеті Proof-Pile-2. Code Llama – модель розроблена Meta для допомоги в написанні, аналізі та оптимізації коду. Вона в свою чергу була ініціалізована з Llama 2 та додатково навчена на 500 мільярдах токенах інформації (за токен приймається будь-яка одиниця тексту, що обробляється моделлю: символ, слово, частинка слова). Proof-Pile-2 – це великий високоякісний набір даних, який включає в себе 55 мільярдів токенів математичних і наукових документів. Він складається із таких датасетів:

- AlgebraicStack (code): це 11 мільярдів токенів вихідного коду, пов'язаного з математикою. Датасет налічує 17 мов серед яких є Lean, Isabelle, Coq (для доведення теорем) та популярні Python, Matlab, C та C++.
- OpenWebMath (web): 15 мільярдів токенів з веб-сторінок, які містять математичний контент
- ArXiv (papers): 29 мільярдів токенів наукових статей.

EleutherAI створили дві версії моделі, що відрізняються кількістю параметрів: llemma_7b та llemma_34b.

Llemma_7b: тренування було здійснене на 200 мільярдах токенах, за 42.000 кроків. Розмір партії становив 4 мільйони токенів і довжиною контексту в 4096 токенів. Це відповідає приблизно 23.000 A100-годинам.

Llemma_34b: тренування було здійснене на 50 мільярдах токенах, за 12.000 кроків. Розмір партії становив 4 мільйони токенів і довжиною контексту в 4096 токенів. Це відповідає приблизно 47.000 A100-годинам.

Оцінювання моделі відбувалося за трьома параметрами: можливість розв'язувати завдання, які потребують вміння послідовного логічного міркування (chain of thought reasoning), використання інструменту few-shot (вміння ідентифікувати тип завдання та вирішити його за шаблоном схожого типу завдань з тренувальних даних) та навички запам'ятовування і змішування даних.

Процес тестування проходив у два етапи: розв'язок задач із точною відповіддю, вирішення задач на доведення.

1) Для аналізу моделі на першому етапі було використано такі датасети:
   - MATH (Hendrycks et al., 2021b) : набір даних з 12,5 тис. завдань (5 тис. оцінок) з математичних олімпіад для старшокласників. Для тестування було застосовано підхід Lewkowycz et al. (2022) 4-shot промптінг, а для перевірки результату Python бібліотеку SymPy, яка може виявити еквівалентні стрічки, якщо вони написані по-різному.
   - GSM8k (Cobbe et al., 2021): набір даних задач з математики для учнів середньої школи.
   - OCWCourses (Lewkowycz et al., 2022): колекція задач для студентів Массачусетського технологічного інституту програми бакалаврату.

- MMLU-STEM (Hendrycks et al., 2021a): підгрупа з 18/57 предметів за критерієм MMLU.
- SAT: створений розробниками датасет із 32 завдань

Для порівняння успішності LLEMMA було обрано модель Minerva, створену Google Research. Ця модель цікава тим, що продовжила навчання PaLM.

Нижче наведені результати тестування, за якими можна зробити висновок, що показники ефективності LLEMMA перевищують досягнення інших моделей з еквівалентною кількістю параметрів.

		GSM8k	OCW	MMLU-STEM	SAT	MATH
Llama 2	7B	11.8%	3.7%	29.9%	25.0%	3.2%
Code Llama	7B	10.5%	4.4%	25.1%	9.4%	4.5%
Minerva	8B	16.2%	7.7%	35.6%	-	14.1%
LLEMMA	7B	**36.4%**	**7.7%**	**37.7%**	**53.1%**	**18.0%**
Code Llama	34B	29.6%	7.0%	40.5%	40.6%	12.2%
LLEMMA	34B	**51.5%**	**11.8%**	**49.0%**	**71.9%**	**25.0%**
Minerva	62B	52.4%	12.0%	53.9%	-	27.6%
Minerva	540B	58.8%	17.6%	63.9%	-	33.6%

2) Для автоматизованого генерування доведень існують такі помічники як Lean (de Moura et al., 2015), Isabelle (Wenzel et al., 2008) та Coq (Paulin-Mohring, 1989a,b), які виражають математичні твердження та логічні конструкції мовами програмування, які уможливлюють їх верифікацію моделями. Часто таким мовам не приділяють достатньо уваги, тому їх відсоток у навчальних датасетах дуже малий. AlgebraicStack Proof-Pile-2 містить понад 1,5 мільярда токенів формальних математичних даних, включаючи дані різних станів доведення, витягнутих з формалізацій Lean та Isabelle.

   Перевірка моделі була проведеня на двух типах задач:

   - Informal-to-formal proving

Завдання моделі полягало у генеруванні формального доведення на основі формального твердження, неформального LATEX-твердження та неформального LATEX доведення, для того щоб надалі результат міг бути перевірений помічником Isabelle Для тестування було обрано 13 задач: 7 з теорії чисел і 6 з інших галузей. Результати перевірки за бенчмарком miniF2F також показали перевагу моделі LLEMMA над іншими.

Method	**Informal-to-formal**	
	miniF2F-valid	miniF2F-test
Sledgehammer	14.72%	20.49%
Code Llama 7b	16.31%	17.62%
Code Llama 34b	18.45%	18.03%
LLEMMA-7b	20.60%	22.13%
LLEMMA-34b	21.03%	21.31%

- Formal-to-formal proving

  На даному етапі тестування завданням моделі було згенерувати доведення в кооперації з помічником Lean 4, а саме писати продовження або наслідки до його тверджень. Після кожного кроку помічник перевіряв вивід моделі і в разі успішного проходження цього кроку, переходив на наступний. Процес виконання зупинявся, коли доведення було завершене або по закінченню встановленого терміна виконання. Перевірка також була проведена на основі бенчмарка miniF2F.

Method	**Formal-to-formal**	
	Search	miniF2F-test
ReProver*	1×64	26.50%
Code Llama 7b	1×32	20.49%
Code Llama 34b	1×32	22.13%
LLEMMA-7b	1×32	26.23%
LLEMMA-34b	1×32	25.82%

Отже, завдяки різноманітності навчальних даних та процесу тренування, модель LLEMMA охоплює широкий спектр напрямків галузі математики і може бути корисною у вирішенні різнопланових математичних задач. Також важливо пам’ятати, що основою досліджуваної моделі є Code Llama, яка спеціалізується на завданнях пов’язаних з кодуванням, тому LLEMMA може зробити свій внесок у вирішення завдань, що поєднують у собі математику та програмування.

ВИСНОВОК

Провівши дослідження наукових джерел було виявлено, що чат-бот – це комп'ютерна програма, яка створена для того, щоб імітувати живу розмову з користувачами. За набором функціональних властивостей чат-боти можна поділити на дві категорії: помічники, які працюють за заданим сценарієм та використовують сталі відповіді та ті, в які був інтегрований ШІ, завдяки чому вони можуть бути більш гнучкими до користувацьких запитів. Наразі особливу увагу приділяють саме другому типу, над яким ведуться подальші дослідження та розробка. Основним завданням, яке ставлять перед собою розробники, є адаптація чат-ботів до стилю мовлення користувача, його настрою та особливостей менталітету. Також існує класифікація за сферою спеціалізації: універсальні чат-боти та орієнтовані на питання лише в певній галузі. Вони різняться набором навчальних даних, тому більш точну відповідь на запитання з конкретної предметної області ймовірніше дасть чат-бот, який на цьому спеціалізується. Головна складність розробки помічника – вести контроль за етикою відповідей ШІ. Через різноманітність ресурсів з яких формуються навчальні датасети, існує доволі висока вірогідність, що в них можуть потрапити матеріали, які не відповідають нормам права та моралі. Саме тому найважливішим етапом розробки моделей є створення правил та норм, яким має слідувати програма.

У наступному розділі першого підрозділу було розглянуто найпопулярніших, станом на сьогоднішній день, представників чат-ботів, які будуть протестовані та проаналізовані у наступному розділі, й використані для подальшого дослідження. Серед обраних помічників ChatGPT, Claude AI, Gemini та Perplexity AI. Перші три створені на основі власних моделей, процес навчання яких та освітні дані суттєво відрізнялись. Perplexity AI поєднує у собі декілька моделей, до яких входять GPT та Claude, проте він був обраний завдяки своїм головним перевагам – потужний пошуковий механізм та доступ до найновіших

актуальних даних. Чат-боти існують у декількох версіях, тому вище зазначено які саме було вибрано для порівняння за ефективністю із досліджуваною моделлю.

Розділ присвячений моделі LLEMMA містить основну інформацію про саму модель, її розробку та тестування. Вона була створена неприбутковою організацією EleutherAI. Розроблена на основі моделі Code Llama та додатково навчена на датасеті Proof-Pile-2, LLEMMA була створена для допомоги у вирішенні математичних завдань. До набору навчальних даних увійшли такі дані як програмний код, інформація з веб-сторінок, які містять математичний контент, та матеріали наукових статей. Існує дві версії моделі, які різняться кількістю параметрів: llemma_7b та llemma_34b. За результатами тестування, у яке були включені як завдання, які вимагають конкретної відповіді, так і задачі на доведення, обидві версії показали найкращі результати серед порівнюваних моделей. Завдяки своїй спеціалізації та успішності серед схожих моделей, LLEMMA стала цікавим об’єктом для тестування, аналізу та подальшого дослідження.

РОЗДІЛ 2: Аналіз моделі llemma_7b

2.1 Головне завдання дослідження та спосіб його проведення

Основною ціллю даного етапу роботи було визначити ефективність моделі llemma_7b та її особливості використання провівши тестування, та якісний і кількісний аналізи на основі отриманих результатів, для того, щоб визначити чи придатна дана версія моделі для реалізації загальної мети по створенню чат-бота для допомоги у вирішенні математичних завдань. Також важливим завданням було протестувати інші загальновідомі чат-боти, які не спеціалізуються на вирішенні завдань з математики, але потенційно можуть бути використані для цього, та провести порівняльний аналіз з показниками досліджуваної моделі.

Для проведення практичної частини даного етапу було створено набір даних, який включає в себе завдання з 11 розділів математики та секцію із задачами підвищеної складності та на доведення. Також у роботі міститься додатковий розділ з прикладами призначеними для виявлення особливостей семантики та формулювання найбільш оптимального запиту для взаємодії з моделлю. Кожна частина, якій відповідає певний розділ математики, сформована з 12 завдань різної складності, які потребують формулювання чіткої короткої відповіді. Останній блок складається з 12 завдань, які розраховані відкриту структуровану змістовну відповідь. 95% запитів сформовані англійською мовою, інші 5% написані мовами Python та Isabelle. Більшість завдань були підібрані або створені на основі курсу математики шкільної програми, олімпіадних програм та навчальних матеріалів для здобувачів вищої освіти.

Тестування було проведене в просторі Google Colab та Jupyter Notebook із використанням середовища потужного локального комп'ютера. Для встановлення моделі llemma_7b було застосовано більше 25 ГБ оперативної пам'яті та 61 ГБ фізичної.

Перевірка результатів була здійснена вручну. Відповіді моделі було розсортовано у чотири категорії: правильні, хибні, частково правильні та «нема відповіді». До частково правильних були віднесені приклади, у розв'язаннях до яких модель правильно визначала тип завдання і пропонувала алгоритм вирішення, але з помилялася в обрахунках або частково розв'язані завдання із перспективним початком.

В залежності від складності завдання у запитах було встановлене обмеження на максимальну кількість символів вихідного результату моделі: 200 – 700 символів. На генерацію відповіді на 200 символів було затрачено близько 4 хвилин, на 700 символів – 13.

Для обрахунку загальної успішності моделі була використана така формула:

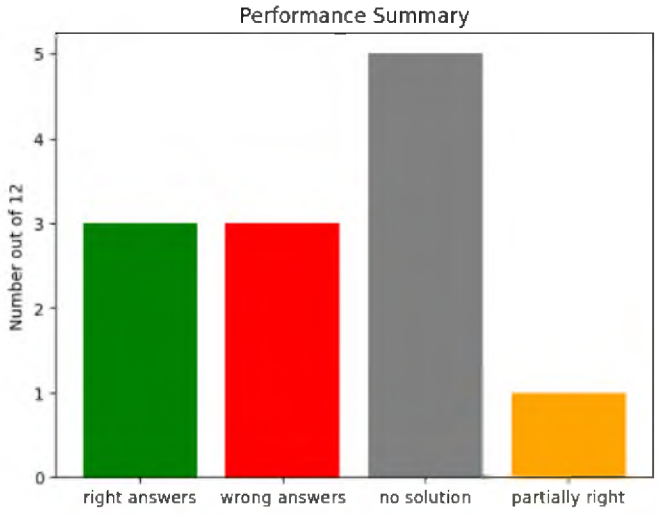
$$SR = \frac{r + \frac{1}{2}p}{a} * 100\%,$$

Де r – кількість правильних відповідей, p – кількість частково правильних відповідей, a – кількість усіх завдань.

2.2 Тестування моделі за розділами математики

1) Арифметика

Цей розділ математики вивчає властивості чисел та дії над ними. Нижче представлена гістограма, яка відображає статистику по даному розділу, а саме кількість правильних відповідей, хибних, нерозв'язаних завдань та частково правильних рішень.



Опис та аналіз результатів:

Модель добре впоралась із операцією додавання на одноцифрових числах, проте результат добутку трицифрових чисел виявився хибним. Наступні 4 приклади були присвячені обрахунку числа в певному степені за модулем: 2 завдання були записані у звичайному форматі $x^y (mod\ z)$, інші два завдання містили значення, ідентичні до перших двох, але були записані мовою Python - $(x ** y)\ \%\ z$. Модель дала правильну відповідь лише на одне завдання задане на Python, в якому фігурували одноцифрові числа, з чого можна зробити висновок, що більш вірогідно отримати правильну відповідь на даний тип завдання надіславши вхідні дані у форматі коду, проте із складними обчисленнями дана версія моделі не справляється.

Два завдання, які містили в собі три операції, включаючи піднесення до степеня та обчислення квадратного кореня, також не були виконані

успішно, проте із завданням на обчислення кубічного кореня модель впоралась. З чого можна зробити висновок, що складність полягала у кількості операцій в одному прикладі.

Факторіал модель правильно розписала, проте не обчислила, тому це завдання було віднесене до групи “partially right”.

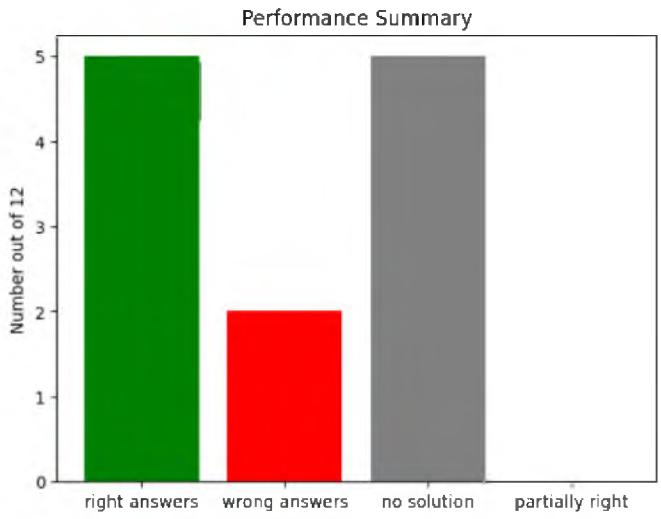
Для завдань з обчислення логарифмів модель не надала розв’язку.

Успішність моделі: 29,17%

2) Алгебра

Цей розділ математики вивчає операції та відношення математичних об’єктів, а також методи розв’язання рівнянь і систем рівнянь.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Досліджувана модель не змогла знайти розв’язок звичайного лінійного рівняння, проте продемонструвала здатність до успішного розв’язання завдань на знаходження значення функції при заданому значенні x.

З трьох квадратних рівнянь правильну відповідь та шлях розв’язання було виявлено у двох результатах.

Із чотирьох завдань з розкладу виразу на множники легкої та середньої складності модель дала коректну відповідь на два.

Також виявилось, що модель не здатна розв’язувати завдання на розкриття дужок та спрощення виразів, так само як і на вирішення системи рівнянь.

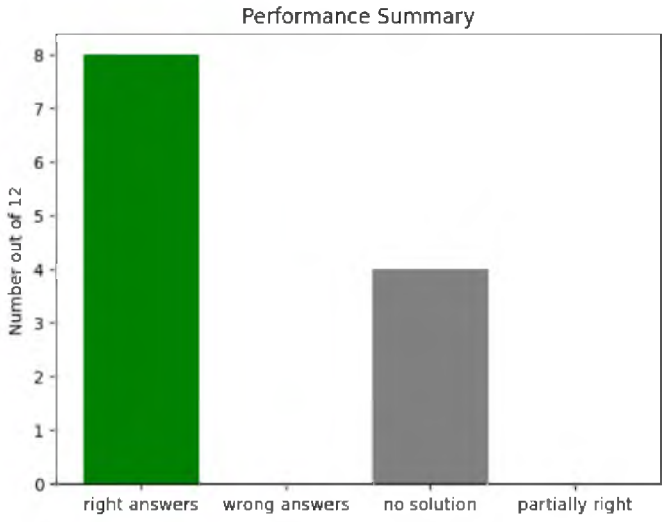
Отже, на основі результатів тестування можна стверджувати, що модель здатна вирішувати завдання, які потребують знання формул та вміння правильно підставити у них значення. Загалом можна зазначити, що хороші результати були отримані у найпростіших екземплярах кожного типу завдань.

Успішність моделі: 41,67%

3) Геометрія

Цей розділ математики вивчає просторові фігури, їх властивості та відношення.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Завдання на обчислення периметра фігур, площі, об’єму куба, довжини кола, за заданими необхідними величинами були виконані успішно. Тобто модель може коректно визначити формулу, підставити у неї дані з умови завдання та обчислити відповідь.

За результатами тестування моделі на завданнях з обчислення об’єму конуса та циліндра можна визначити, що відповіді моделі результативні тільки у випадках, коли вхідні числові значення невеликі.

Також було виявлено, що модель вміє підставляти необхідні значення у формулу, проте якщо завдання потребує кількох кроків розв’язання, вона не може вважатись ефективною.

Досліджуючи результати виконання завдань на вектори виявилось, що модель ефективна у визначенні колінеарності, проте не здатна обчислити довжину вектора.

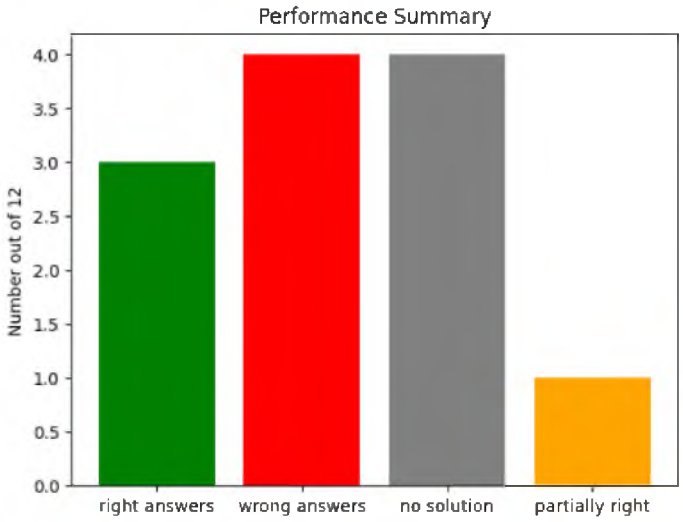
Із записом рівняння кола з центром у точці початку координат, яке проходить через точку із заданими координатами, модель також не впоралась.

Успішність моделі: 66,67%

4) Тригонометрія

Цей розділ математики вивчає відношення між сторонами та кутами трикутників, а також тригонометричні функції та їх застосування.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Завдання на визначення косинуса певного значення у радіанах та тангенса кута у градусній мірі було представлено у кількох прикладах з різним

формулюванням, але усі завдання модель не виконала або надала хибну відповідь. В завданні на обрахунок синуса модель допустила помилку, проте зазначила правильну відповідь у коментарях до свого розв'язку, тому цей розв'язок був віднесений до «частково правильних» відповідей. Також модель помилилася у визначенні тангенса кута більше 90°, коли у формулюванні фігурували символи «°», «'», проте дала правильну відповідь, коли було надіслано запит, що містив слово «radians».

Незважаючи на помилки у найпростіших завданнях, модель правильно обрахувала арксинус та знайшла sin(2x) знаючи sin(x), з чого слідує, що модель ефективна у вирішенні завдань на знання формул, але не завжди справляється із завданнями на обчислення.

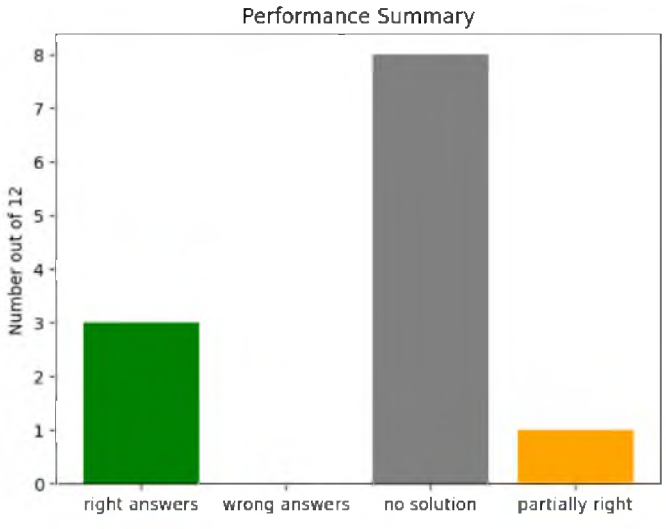
Також у ході тестування було виявлено, що модель не вміє вирішувати тригонометричні рівняння, що вважаються завданнями середньої складності.

Успішність моделі: 29,17%

5) Математичний аналіз

Цей розділ математики вивчає границі послідовностей та функцій, диференціювання (похідні) та інтегрування функцій однієї та кількох змінних.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Усі завдання на знаходження похідної та другої похідної були розв'язані правильно, у той час як всі завдання на визначення інтеграла (з межами, без меж, задача на геометричний зміст, на обчислення за вказаною формулою) – неправильно.

Із завданнями на знаходження критичних точок, лімітів, максимального значення функції модель також не впоралась.

Ускладнене завдання на обчислення значення похідної в певній точці модель змогла вирішити після надсилання покращеного запиту з інструкціями щодо послідовності розв'язання.

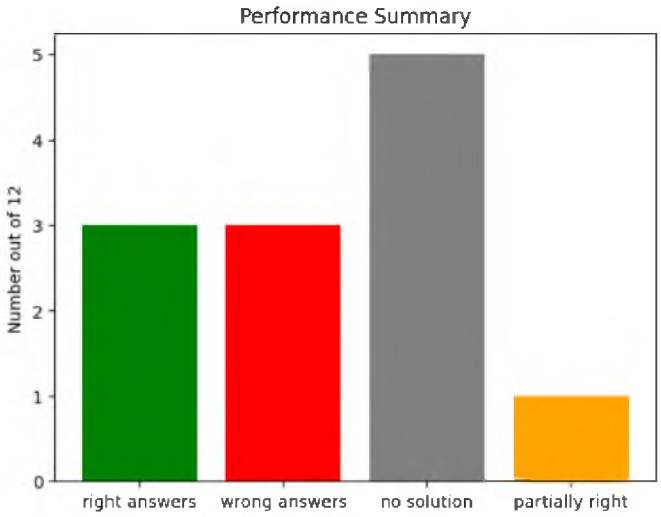
Отже, модель може стати у асистентом під час вирішення задач орієнтованих на перевірку знань у темі похідної, проте в інших завданнях цієї галузі математики вона не вважається ефективною.

Успішність моделі: 29,17%

6) Теорія вірогідності

Цей розділ математики вивчає випадкові події та величини, їх властивості та операції, які можна над ними здійснювати.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Дані для тестування налічували 3 завдання на підкидання кубика із 2 з яких, модель впоралась успішно, проте жодна із задач на підкидання монети не була розв'язана правильно.

Із 4 завдань на вірогідність вибору певного елемента із множини всіх елементів одне було згенероване за допомогою ChatGPT, саме воно було розв'язане і пояснене правильно, проте на три інших завдання цього ж типу був наданий хибний розв'язок. Після перегляду відповідей цього блоку виникло припущення, модель з більшою вірогідністю правильно розв'яже завдання сформульовані у контексті тематики книг або кульок. Тому було переписано задачу про вірогідність вибору певного виду меду на задачу про кульки, зберігаючи усі числові значення та особливості формулювання завдання. Припущення справдилось: модель впоралась із другим варіантом задачі, з чого можна зробити висновок, що вона вміє правильно виявити тип завдання та визначити необхідний для його розв'язання алгоритм, проте тільки тоді, коли завдання сформульоване у контексті найбільш розповсюджених тематик для даного типу зачач.

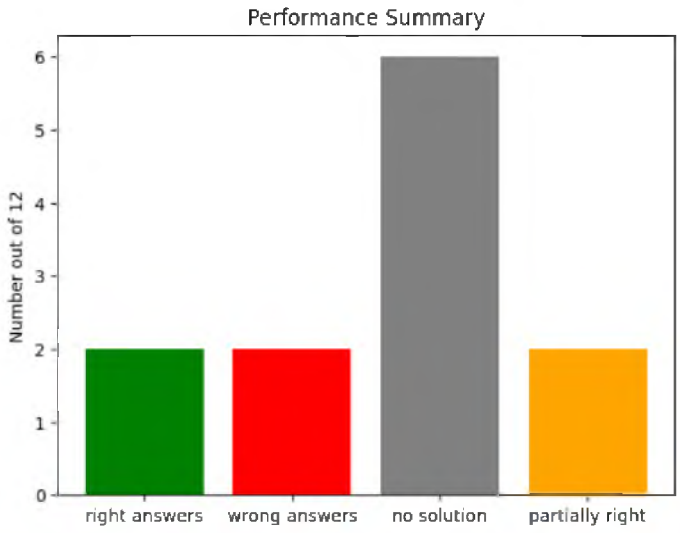
У відповіді до задачі на умовну ймовірність модель правильно зазначила формулу та інтерпретувала її, проте неправильно визначила ймовірності, тому це завдання було зараховане як частково правильне.

Успішність моделі: 29,17%

7) Статистика

Цей розділ математики вивчає організацію, аналіз та інтерпретування даних.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Завдання, подані у декількох варіантах формулювання, на обчислення середнього значення вибірки, медіани та середнього квадратичного відхилення не були правильно інтерпретовані і розв'язані, проте модель дала коректні відповіді до завдань на пошук моди числового ряду та розмаху вибірки.

Замість міжквартильного розмаху, модель рахувала звичайний розмах вибірки, тому шляхом переформулювання запитання було визначено, що модель знає це поняття і може дати йому визначення («The interquartile range is the difference between the upper quartile and the lower quartile»), але воно недостатньо точне для того, щоб правильно вирішити завдання.

Коефіцієнт кореляції модель визначила неправильно через помилку в обчисленні середнього квадратичного відхилення, проте коваріацію було обчислено коректно, завдяки чому розв'язок був оцінений як «частково правильний».

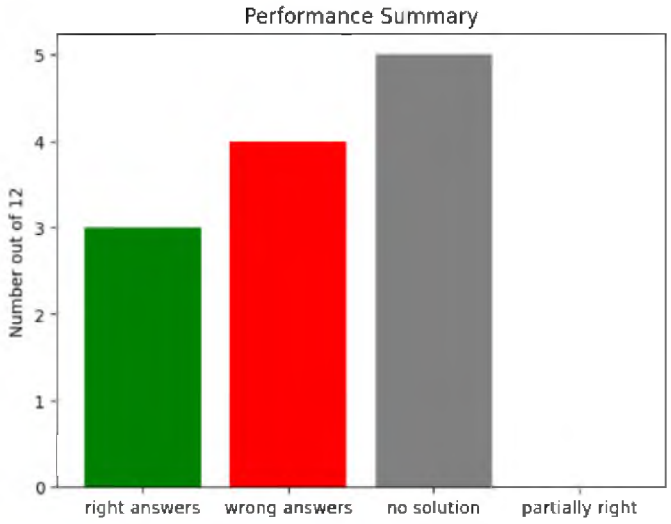
Прикладну задачу модель не змогла інтерпретувати, з чого можна зробити висновок, що найоптимальніший спосіб постановки задачі – запис у формі числових рядів без додаткової інформації про контекст задачі.

Успішність моделі: 25%

8) Теорія чисел

Цей розділ математики присвячений переважно вивченню цілих чисел та арифметичних функцій.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Завдання на визначення найбільшого спільного дільника (НСД) та найменшого спільного кратного (НСК) двоцифрових чисел модель виконала правильно, але визначити чи є число простим або знайти усі прості дільники числа вона не змогла.

Вирішення конгруентностей та обчислення функції Ейлера також не були вирішені коректно.

Визначити чи є число паліндромом модель змогла правильно, додавши пояснення до розв’язку.

Задачі на спрощення виразу, арифметичну та геометричну прогресії не були інтерпретовані моделлю.

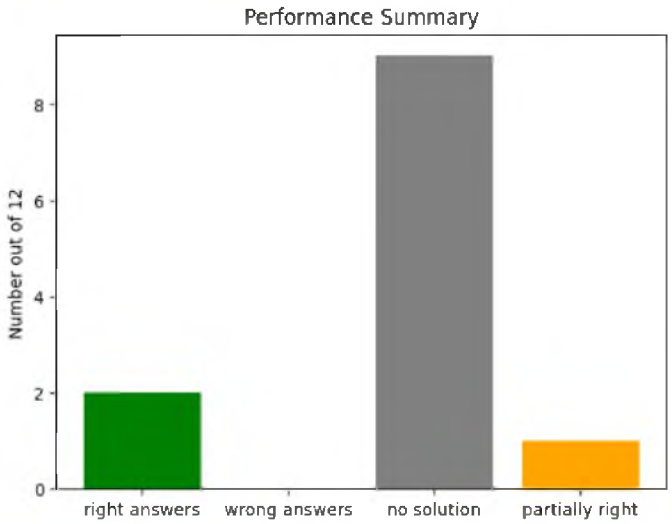
Отже, модель може вважатись ефективною для розв’язання лише найпростіших завдань даного розділу математики.

Успішність моделі: 25%

9) Диференціальні рівняння

Цей розділ математики вивчає рівняння, які пов'язують одну або декілька невідомих функцій та їхні похідні.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

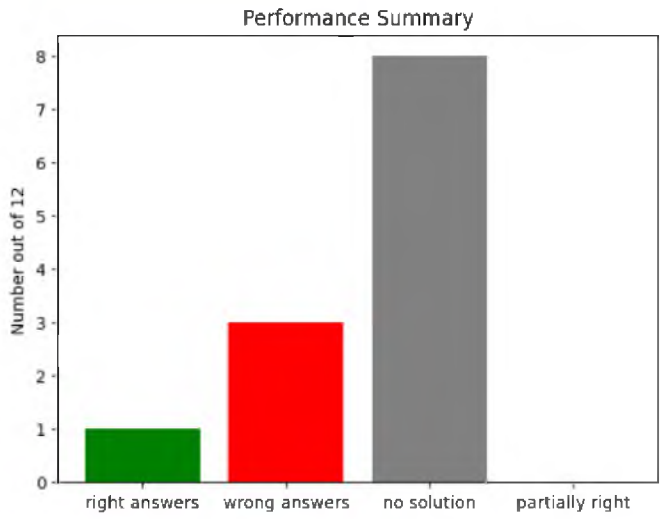
Модель успішно впоралась із розв’язанням диференціального рівняння першого порядку із заданою початковою умовою та із записом загального розв’язку для диференціального рівняння другого порядку. Складність завдань була невеликою, проте дані приклади демонструють наявність у моделі навичок для розв’язання задач вищої математики. Проте на трошки складніші рівняння модель вже не дала відповіді. Також варто зазначити, що перші два приклади були згенеровані за допомогою ChatGPT, а умова до жодного з інших рівнянь, до написання якої не була залучена жодна з нейронних мереж, не була інтерпретована правильно. Частково правильно було розв’язане завдання з неоднорідним диференціальним рівнянням, де була виявлена помилка в одному з коефіцієнтів.

Успішність моделі: 20,83%

10) Лінійна алгебра

Цей розділ алгебри вивчає вектори, векторні простори, лінійні відображення та системи лінійних рівнянь.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Завдання на пошук власних чисел та власних векторів, пошук рангу матриці, детермінанту в різних формулюваннях не були інтерпретовані моделлю, єдиним винятком став варіант формулювання «what is determinant of …», який модель змогла визначити, але дала хибну відповідь.

Правильна відповідь була отримана у завданні на додавання матриць 2×2, а от рівняння із множенням матриць тої самої розмірності, модель не змогла розв’язати коректно.

Лінійні рівняння також не були вирішені навіть частково, як і завдання на пошук координат вектора в натуральному базисі.

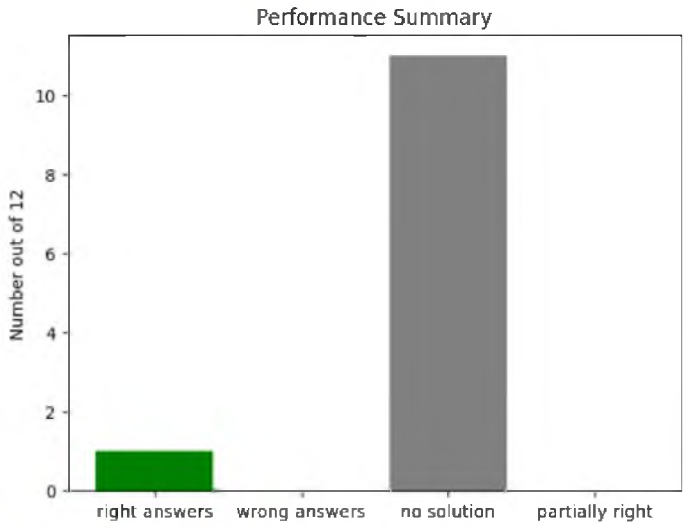
Таким чином можна припустити, що в навчальних даних моделі завдань з лінійної алгебри було недостатньо для того, щоб вона могла правильно навіть найпростіші завдання основних типів задач лінійної алгебри.

Успішність моделі: 8,33%

11) Комбінаторика

Цей розділ математики займається проблемами вибору, розташування та функціонування в межах скінченної або дискретної системи.

Статистика відповідей даного розділу:



Опис та аналіз результатів:

Правильно була розв'язана лише одна задача про кількість варіантів розстановки книг, що демонструє схожу до висновку в розділі «Теорія ймовірності» ідею: модель може коректно розпізнати завдання, основною тематикою яких є книги або кульки.

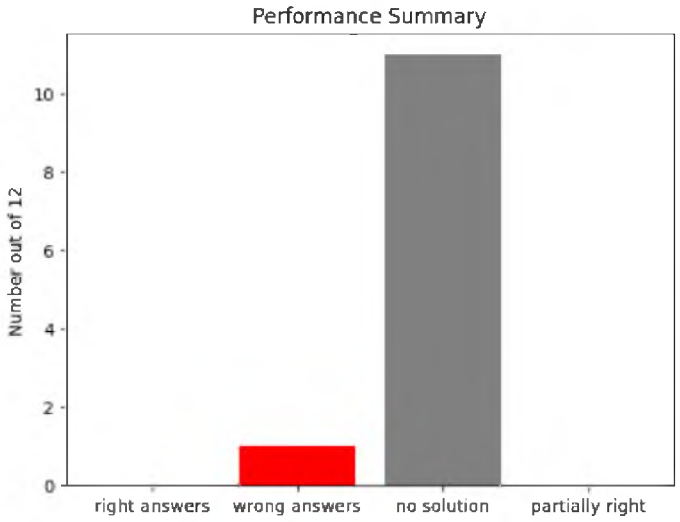
Інші задачі на перестановки, розміщення та комбінації не були правильно інтерпретовані, що свідчить про те, що модель не пристосована для ефективного вирішення текстових задач, через складність виявлення відповідних патернів у неструктурованій формі подання умови.

Успішність моделі: 8,33%

12) Задачі підвищеної складності та на доведення

Цей розділ містить в собі завдання, які потребують багатокрокового розв'язання, відкритої відповіді або чіткого доведення. Задачі останнього типу були сформульовані двома мовами: англійською та Isabelle.

Статистика відповідей даного розділу:



Задача на доведення неможливості трьох чисел бути послідовними була єдиною, яку модель правильно інтерпретувала, проте надала неправильний розв'язок.

Задачі на доведення найпопулярніших теорем (т.Піфагора, т.Ферма) не були розпізнані моделлю, як і усі інші завдання підвищеної складності. Незважаючи на те, що у моделі виникають складнощі у розумінні умов сформульованих текстом, вона дає відповідь на питання про загальні алгоритми розв'язання певних типів задач. Також зміна формулювання завдань має вплив на якість відповіді моделі. Наприклад, задачу у формулюванні «визнач чи дане твердження X правдиве або хибне» модель розтлумачити не змогла, проте на запит «твердження X правдиве або хибне?» була надана правильна відповідь.

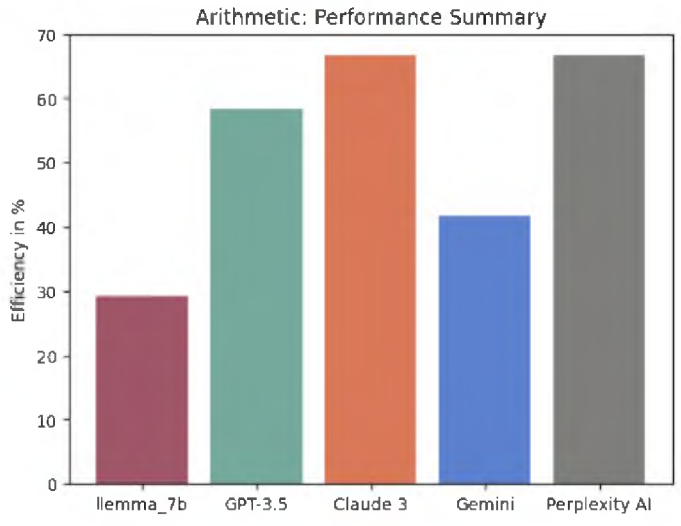
Успішність моделі: 0%

2.3 Тестування обраних чат-ботів та порівняння з ефективністю досліджуваної моделі

В даному розділі буде проведене тестування моделей GPT-3.5, Claude 3 Sonnet, Gemini, Perplexity AI на стеку завдань представленого у розділі 2.2 та буде здійснено порівняльний аналіз за ефективністю чат-ботів та моделі llemma_7b.

1) **Арифметика**

Статистика ефективності моделей:



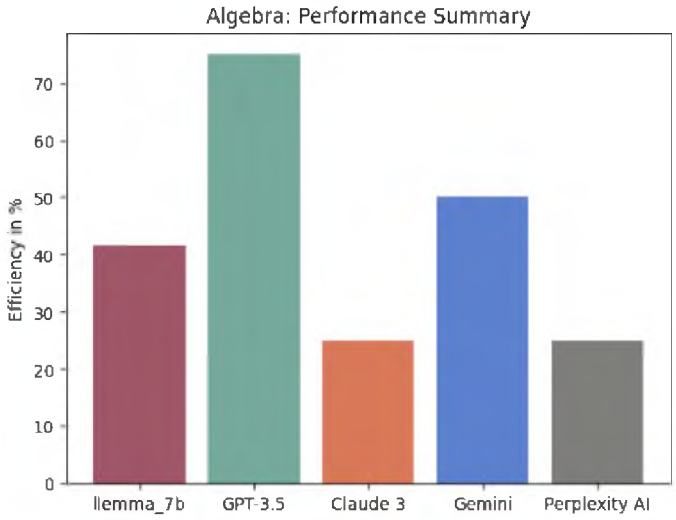
Опис та аналіз результатів:

Найбільш успішними у розв'язанні завдань даного розділу математики виявились моделі Claude 3 та Perplexity AI. На одну правильну відповідь менше дав ChatGPT помилившись в обчисленні добутку трицифрових чисел.

Найрозповсюдженішими помилками були неточності в обчисленні числа в степені за певним модулем та виразів з кількома операціями та коренем числа. Найбільш близькими до правильної відповіді були Claude 3 та Perplexity AI.

2) Алгебра

Статистика ефективності моделей:



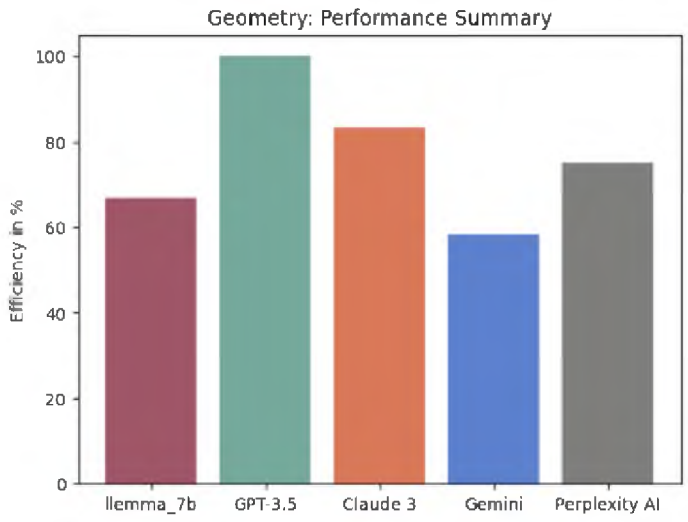
Опис та аналіз результатів:

Модель GPT виявилась найефективнішою у вирішенні алгебраїчних завдань. Найбільш розповсюдженою була помилка у завданні на розкриття дужок та спрощення виразу. Воно було подане у двох формулюваннях, для одного з яких ChatGPT дав правильну відповідь. Модель llemma_7b показала коректний результат для ще одного завдання на спрощення, на відміну від моделей Claude, Gemini, Perplexity, які не впорались із цим завданням. Також вона знайшла правильні дробові корені квадратного рівняння, для якого жодна із популярних моделей не дала правильної відповіді.

Також у ході тестування була виявлена схожість у хибних відповідях моделей Claude 3 та Perplexity. Згідно з документацією Perplexity AI у своїй звичайній версії має використовувати ChatGPT, як додаткову інтегровану модель, проте результати дослідження свідчать про те, що для проведення обрахунків була використана модель Claude.

3) Геометрія

Статистика ефективності моделей:

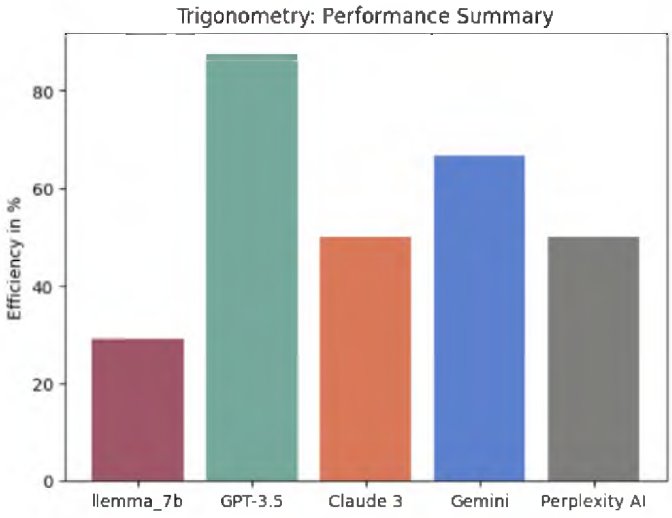


Опис та аналіз результатів:

100% правильних відповідей на завдання даного розділу дала модель GPT-3.5. Найбільш складним завданням для моделей виявилось дослідити вектори на колінеарність, з яким впоралась тільки моделі llemma_7b та GPT. Також при обчисленні об'єма конуса були допущенні помилки, але для такого ж завдання, але з меншими числами, були дана правильна відповідь усіма моделями, що підтверджує те, що складність виникає саме в процесі обчислення, а не пошуку відповідних формул та підходів для розв'язання завдання.

4) Тригонометрія

Статистика ефективності моделей:

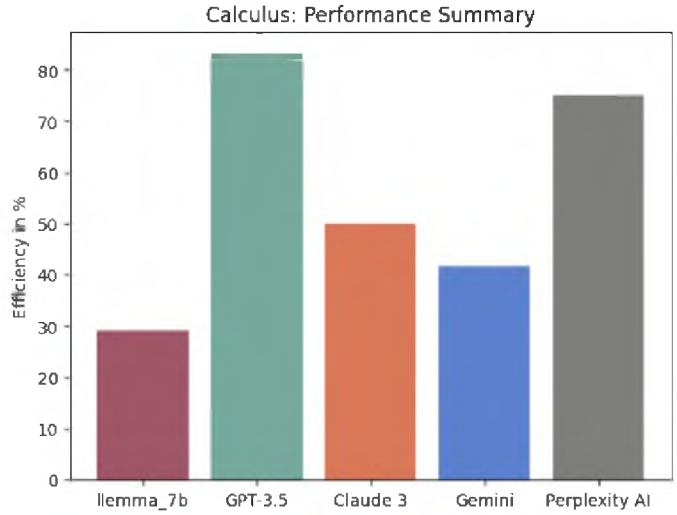


Опис та аналіз результатів:

Найбільше помилок було зроблено у завданнях на пошук косинуса та синуса подвійного кута, обчислення тангенса кута більше 90 градусів та на розв'язання тригонометричного рівняння. Llemma_7b успішно впоралась із обчисленням синуса подвійного кута, проте дала хибну відповідь у завданнях на відображення знань метрик табличних кутів. На тригонометричне рівняння була надана лише одна правильна, проте неповна відповідь, яку запропонував ChatGPT.

5) Математичний аналіз

Статистика ефективності моделей:

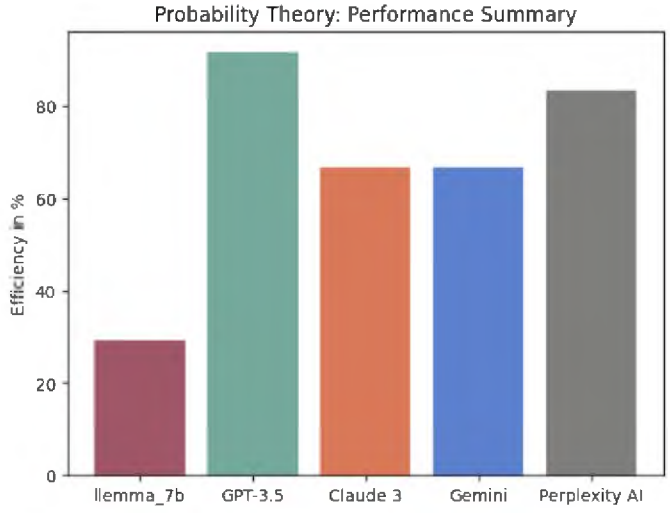


Опис та аналіз результатів:

До списку найважчих для вирішення завдань увійшли обчислення інтеграла з межами та знаходження максимального значення функції, з якими не впоралась жодна з моделей. У другому завданні проблема полягала саме в обчисленні значення функції після підстановки правильно визначеного «x». Наступним по складності виконання виявилось завдання на пошук значення похідної функції у певній точці, з яким впорались лише llemma_7b та GPT.

6) Теорія ймовірності

Статистика ефективності моделей:



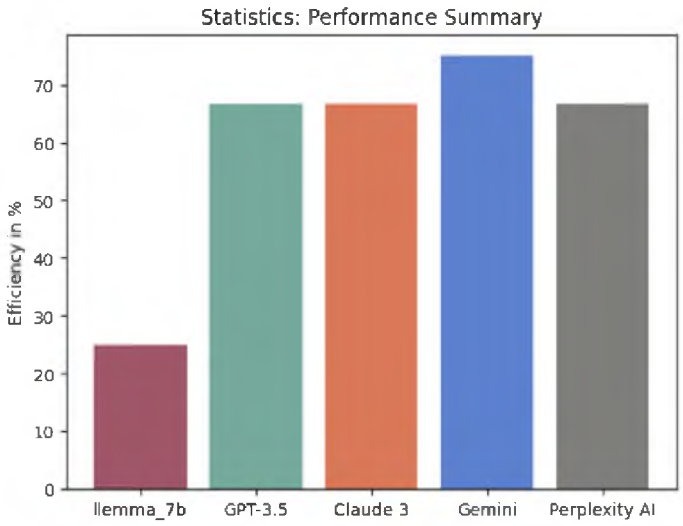
У даному розділі обрані для порівняння моделі демонструють наявність навичок інтерпретації умов задач різних формулювань і тематик, у той час як llemma_7b може розпізнати лише завдання написані у найбільш розповсюджених формулюваннях.

Задачу на умовну ймовірність не вирішила жодна з моделей. Також складною виявилась задача на вірогідність здійснення подій у певній послідовності, яку вирішила лише модель GPT.

Окрім ChatGPT доволі високу ефективність продемонструвала модель Perplexity AI, яка допустила помилки лише у двох завданнях з 12.

7) Статистика

Статистика ефективності моделей:

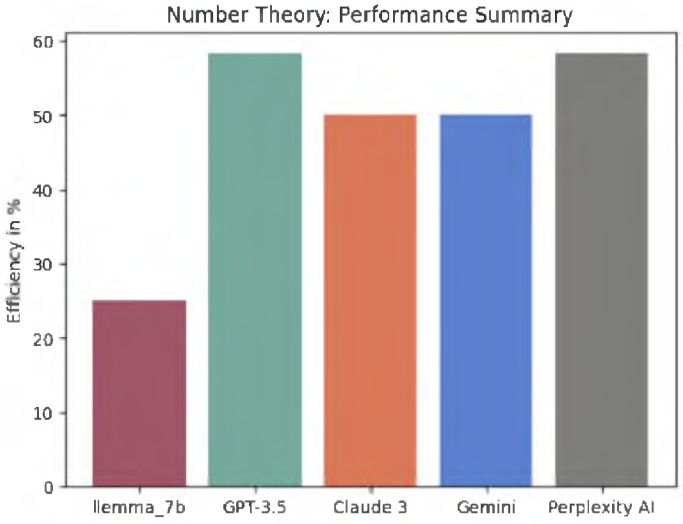


Опис та аналіз результатів:

Одними з найважчих завдань стали розрахунок міжквартильного розмаху, яку правильно вирішила лише модель Gemini, та задача на знаходження невідомого «x» на основі інших даних та середнього значення вибірки, на яке усі моделі дали хибну відповідь. Результати завдання на обчислення середнього квадратичного відхилення трохи варіювалися, але у межах відхилення (+ – 0.5).

8) Теорія чисел

Статистика ефективності моделей:



Опис та аналіз результатів:

Однаково успішними у виконанні завдань цього розділу стали моделі GPT та Perplexity AI, давши правильну відповідь на 7 завдань із 12.

Завдання на розв'язання рівняння-конгруенції правильно виконала лише модель Perplexity AI, проте допустила помилку у визначенні чи є певне число простим.

Завдання високого рівня складності не розв'язала жодна з моделей.

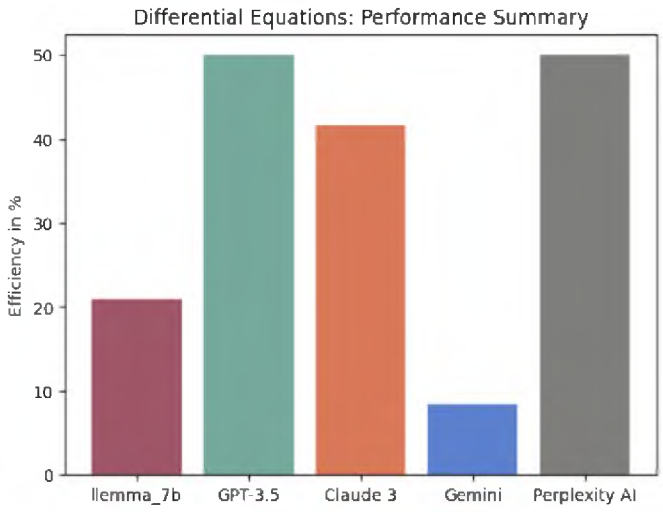
Також серед прикладів для тестування було завдання на арифметичну прогресію із неповною умовою, в якій було недостатньо даних для розв'язання, проте жодна з моделей цього не вказала.

Правильну відповідь до завдання на геометричну прогресію дала лише модель GPT.

Отже, у ході тестування було виявлено, що жодна з моделей не здатна вирішувати завдання, які потребують логічного мислення, багатокрокового розв'язку та робити перевірку своїх відповідей згідно з умовою завдання, перед надсиланням відповіді.

9) Диференціальні рівняння

Статистика ефективності моделей:



Опис та аналіз результатів:

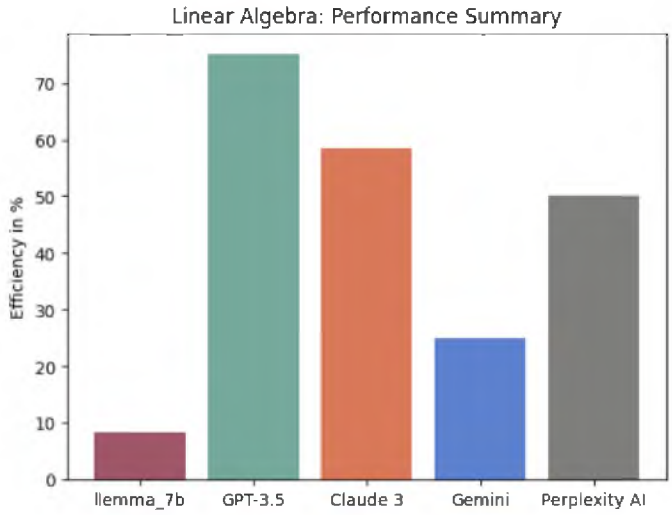
Усі моделі, окрім Gemini, можуть вважатися ефективними для розв'язання найпростіших диференціальних рівнянь, проте завдання із невеликим

ускладненням моделі правильно вирішити не можуть. Єдине завдання, з яким впоралися усі моделі – це звичайне диференціальне рівняння першого порядку.

Отже, за результатами тестування: Gemini – найменш дієва модель для розв'язання диференціальних рівнянь, найкращі показники демонструють моделі GPT та Perplexity AI, що становить 50% правильних відповідей на запропоновані задачі.

10) Лінійна алгебра

Статистика ефективності моделей:



Опис та аналіз результатів:

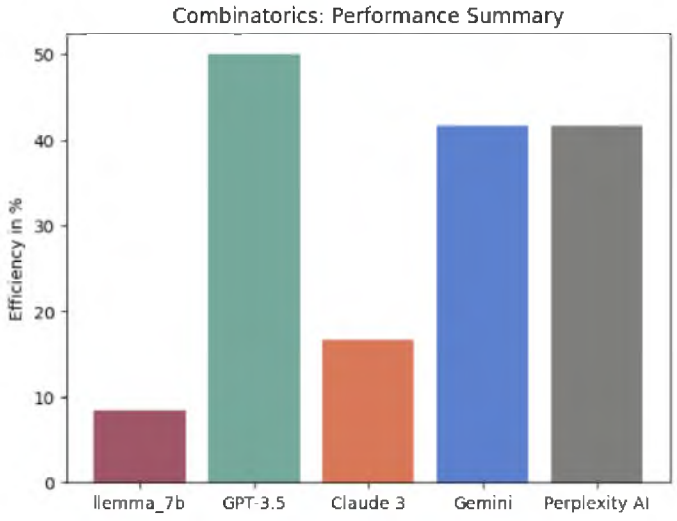
Серед завдань, на які жодна з моделей не дала правильну відповідь, були визначення оберненої матриці, розв'язання матричного рівняння та обчислення визначника матриці 3×3, що підтверджує низьку ефективність моделей у завданнях, де основою розв'язку постають обчислення.

Систему лінійних рівнянь було розв'язано лише програмою ChatGPT.

Незважаючи на те, що у Gemini загалом низькі показники для даного розділу, модель правильно вирішила завдання на пошук власних чисел та власних векторів, яке було правильно розв'язане лише моделлю GPT.

11) Комбінаторика

Статистика ефективності моделей:



Опис та аналіз результатів:

Найбільше помилок було допущено у завданнях, які потребують важких обчислень.

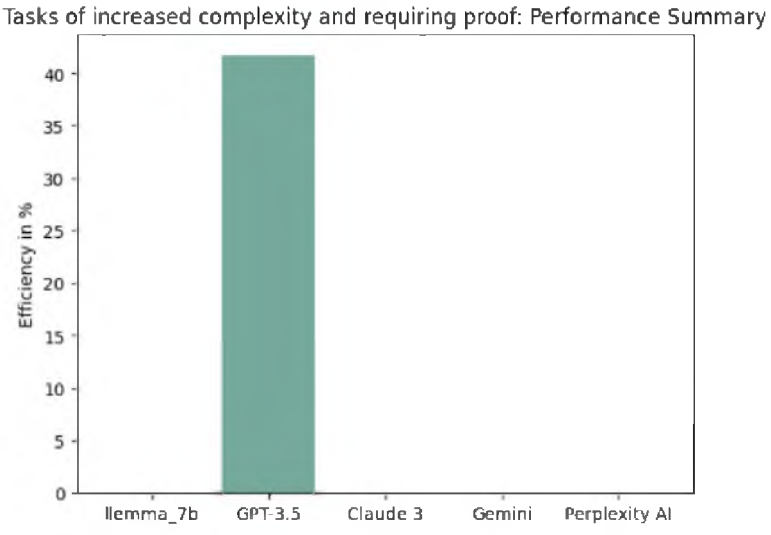
Задача про кількість можливих варіантів здійснення події за певної умови було правильно розв'язане лише моделлю Perplexity AI, а задачу про кількість варіантів утворення правильних дробів з чисел – Gemini.

Задачу із унікальним формулюванням, змогла правильно визначити та розв'язати лише модель GPT.

Отже, найбільш ефективною у вирішенні завдань з комбінаторики серед порівнюваних моделей можна вважати GPT-3.5. Головною складністю завдань цієї галузі математики виступають обчислення, та важливість найменших деталей. Більшість відповідей відрізнялись від правильної у два рази, що свідчить про те, що основну частину роботи моделі виконують правильно, проте роблять помилку не враховуючи певні особливі випадки.

12) Завдання підвищеної складності та на доведення

Статистика ефективності моделей:



Опис та аналіз результатів:

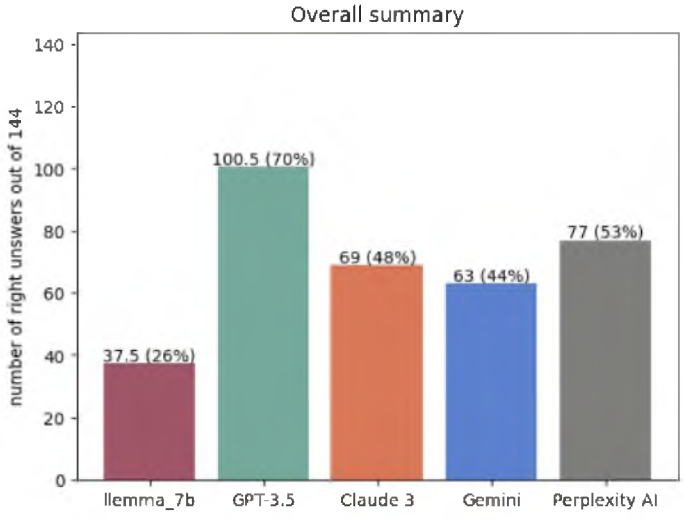
У даному розділі лише одна модель продемонструвала здатність точно обґрунтовувати свої розв'язки та писати доведення найпростіших теорем. ChatGPT дав достатньо повну логічну правильну відповідь на 5 завдань із 7. Також перевагою цієї моделі стало знання мови Isabelle для формування запитів на доведення теорем.

Claude 3 не написала жодного розв'язку, додавши що не має достатньо можливостей для вирішення завдань такого типу.

Gemini та Perplexity AI розв'язали декілька завдань, проте надали хибну відповідь, а для доведення теорем надіслали тільки можливі шляхи розв'язання, проте не провели його.

Отже, на даному етапі розробки досліджувані версії моделей не мають достатньо навичок для вирішення завдань підвищеної складності. GPT може вважатись ефективною у порівнянні з іншими моделями.

Підсумки по дослідженню:



Розділ	К-сть правильних відповідей	К-сть неправильних відповідей	Частково правильні відповіді	Без розв'язку	Успішність
Арифметика	3	3	1	5	29,17%
Алгебра	5	2	0	5	41,67%
Геометрія	8	0	0	4	66,67%
Тригонометрія	3	4	1	4	29,17%
Мат. аналіз	3	0	1	8	29,17%
Теорія ймовірності	3	3	1	5	29,17%
Статистика	2	2	2	6	25%
Теорія чисел	3	4	0	5	25%
Диф. рівняння	2	0	1	9	20,83%
Лінійна алгебра	1	3	0	8	8,33%
Комбінаторика	1	0	0	11	8,33%
Завдання підвищеної складності та на доведення	0	1	0	11	0%

Згідно з результатами проведеного тестування найбільш ефективним помічником для вирішення завдань різних розділів математики можна вважати ChatGPT, який дав правильні відповіді на 70% завдань.

Визначаючи ефективність моделей також варто брати до уваги час затрачений на генерацію відповіді. Модель llemma_7b в середньому затрачає на 7-8 хвилин більше часу у порівнянні з іншими моделями, які дають відповідь майже миттєво. Незважаючи на найнижчі показники ефективності серед усіх порівнюваних моделей, слід зазначити, що модель хоч і вузькоспеціалізована, проте має значно менший обсяг, а саме 7 мільярдів параметрів, у порівнянні з 175 мільярдами параметрів моделі GPT-3. Тому ефективність, що становить 26 відсотків, вважається задовільною і представляє схожі результати до заявлених розробниками значень (середня очікувана ефективність – 30.58%).

ВИСНОВОК

Під час проведення якісного аналізу моделі llemma_7b було виявлено, що успішність моделі залежить від форми подання завдання лише у поодиноких випадках. Єдиною закономірністю, яку точно можна прослідкувати, стала залежність відповіді від контексту задач розділів теорії ймовірності та комбінаторики: модель не може правильно інтерпретувати умову з малопоширеним нетиповим контекстом, проте якщо переписати задачу в контексті книжок або кульок, існує велика вірогідність отримати правильну відповідь.

Для проведення кількісного аналізу було створено набір даних із 144 завдань різної складності, з яких модель правильно дала відповідь на 37.5, що становить 26%. Загальна тенденція, що була виявлена в ході дослідження, полягає у тому, що модель не може впоратись із завданнями, що потребують багатокрокового розв'язку або обчислень, в яких фігурують двоцифрові числа і більше. Ефективність моделі збільшується при наданні додаткових вказівок щодо підходу до вирішення завдань. Llemma_7b показала здатність розв'язати найпростіші завдання з усіх запропонованих розділів математики. Найкращі результати: 41,67% і 66,67% були продемонстровані у розв'язанні завдань з алгебри та геометрії відповідно, з чого слідує, що найбільш доцільною сферою застосування моделі є надання допомоги у засвоєнні навчального матеріалу з курсу математики шкільної програми.

Серед усіх характеристик LLM, виявлених під час спостережень, можна визначити основні: значна кількість помилок в обчисленнях, труднощі у розв'язані завдань, що потребують логічного мислення, багатокрокового процесу вирішення, а також завдань на доведення. Найкращі показники серед досліджуваних моделей були досягнуті моделлю GPT-3.5, що надала правильний розв'язок до 100.5 завдань (70%), в той час як кількість правильних відповідей інших моделей лежала у межах від 63 (Gemini) до 77 (Perplexity).

Цікавим спостереженням є те, що модель Gemini, яка загалом показала найнижчу ефективність порівняно з іншими трьома аналізованими LLM, показала найкращі показники у розділі «Статистика». Загалом можна зробити висновок, що попри те, що досліджувані LLM не спеціалізуються на вирішенні математичних завдань, вони можуть бути використані для допомоги у розв'язанні базових завдань кожної галузі математики.

Зважаючи на розмір моделі llemma_7b, що становить 7 мільярдів параметрів, можна стверджувати, що модель доволі перспективна і демонструє непогані результати у порівнянні з великими моделями (GPT-3.5 налічує 175 мільярдів параметрів).

РОЗДІЛ 3: Розробка чат-бота

3.1 Обґрунтування вибору інструментів розробки

Для реалізації backend-частини проекту було обрано мову програмування Python та фреймворк Flask. Такий набір інструментів забезпечує функціональність для виконання поставлених завдань та відкриває можливості для подальшого вдосконалення програми. Python має велику систему бібліотек та фреймворків, розроблених для обробки природньої мови, які пропонують реалізацію таких функцій як попередня обробка тексту, токенізація, тегування частин мови та розпізнавання іменованих сутностей. Також ця мова програмування широко застосовується в галузі машинного навчання, що є великою перевагою під час розробки чат-бота. Завдяки фреймворку Flask у проекті було реалізовано RESTful API, що полегшує взаємодію із frontend-частиною. Цей фреймворк має широкий спектр розширень, що стане в нагоді під час реалізації додаткової функціональності чат-бота.

Frontend-частина проекту була написана мовами JavaScript, HTML та CSS. Цей стандартний набір забезпечує інтерактивність, динамічність та зручність веб-інтерфейсу, а також високий рівень сумісності із сучасними веб-браузерами.

Отже, такий вибір інструментів розробки уможливлює створення інтерактивного функціонального чат-бота, а також розгортання проекту на різних платформах та обслуговування у різних середовищах.

3.2 Основні етапи розробки та реалізовані функції

Процес розробки застосунку можна умовно поділити на 3 етапи:

- Створення інтерфейсу програми
- Імплементація основних функцій та налаштування зв'язку між frontend та backend частинами
- Підключення моделі та обробка її відповідей

В ході виконання першого етапу було створено веб-сторінку, з якою взаємодіятиме користувач. Основними кольорами сайту було обрано білий, персиковий та червоний, що має заохочувати людей взаємодіяти із застосунком, адже ці кольори асоціюються із оптимізмом та комфортом.

Головною функцією, реалізованою під час другого етапу розробки, стало зчитування відповіді користувача, відправлення її для обробки на backend-частині та відображення отриманої відповіді в чаті. Для цього було створено ендпоінт, ініціалізація якого була здійснена за допомогою фреймворка Flask.

Для створення методу обробки відповідей моделі на третьому етапі розробки було використано спостереження з другого розділу роботи. Відповідь моделі включає не тільки розв'язок завдання, а ще й коментарі з веб-сторінок, додаткові схожі завдання або просто текст, що не має змісту. Було виявлено певні патерни серед відповідей моделі, за якими було створено метод, що виділяє серед даних лише ті рядки, які формують точну відповідь на поставлене запитання чи завдання.

ВИСНОВОК

В ході виконання роботи було створено чат-бот на основі моделі llemma_7b (Llemma_chatbot). Було розроблено дизайн застосунку та реалізовано основну функціональність програми, включаючи метод для обробки відповідей моделі, що базується на спостереженнях, здійснених під час експерименту.

Завдяки обраним інструментам розробки, існує можливість подальшого вдосконалення чат-бота та розширення його функціоналу.

Llemma_chatbot є втіленням поставленого завдання зі створення помічника для вирішення стандартних математичних завдань шкільної програми.

ВИСНОВОК

У ході виконання роботи було проведене дослідження загальних відомостей про модель llemma та популярні LLM, такі як: GPT, Claude, Perplexity AI та Gemini.

У рамках другого розділу було створено набір даних для комплексного аналізу моделі, що включає в себе 144 основних завдання згрупованих за розділами математики та додаткові запити для виявлення семантичних особливостей моделі llemma_7b. На сформованому датасеті було проведене тестування моделі, на основі якого було здійснено якісний та кількісний аналізи. На зазначеному наборі даних було також здійснене тестування вищеперерахованих LLM і проведено порівняльний аналіз із спеціалізованою на вирішенні математичних завдань моделлю llemma_7b. Завдяки проведеному дослідженню було визначено оптимальну сферу застосування чат-боту на основі обраної моделі.

Зважаючи на попередньо визначені цілі, було створено чат-бот для допомоги у вирішенні математичних завдань із реалізацією базового функціоналу. Обрані інструменти розробки відкривають можливість для модифікації та подальшого вдосконалення програми.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1) https://research.aimultiple.com/chatbot-applications/
2) https://uk.wikipedia.org/wiki/ChatGPT
3) https://openai.com/blog/chatgpt
4) https://tedai-sanfrancisco.ted.com/glossary/claude/
5) https://theanilbajar.medium.com/all-about-gemini-models-and-training-process-989fc3e25602
6) https://zapier.com/blog/perplexity-ai/
7) https://en.wikipedia.org/wiki/Perplexity.ai
8) https://www.eleuther.ai/releases
9) https://en.wikipedia.org/wiki/EleutherAI
10) https://mathai2023.github.io/papers/45.pdf