

Національний університет «Києво-Могилянська Академія»
Кафедра інформатики факультету інформатики

Точне налаштування LLM для обраних предметних областей

Науковий керівник:
ст. викл. кафедри інформатики
Андрощук М. В.

Підготував:
студент КН-БП-4
Остролуцький А. Б.

Мета роботи

здійснити точне налаштування великих мовних моделей (LLM) для застосування в конкретних предметних областях, а саме:

- науково-теоретичній медицині,
- клінічній медицині,
- фінансовому аналізу.

Актуальність і проблематика дослідження

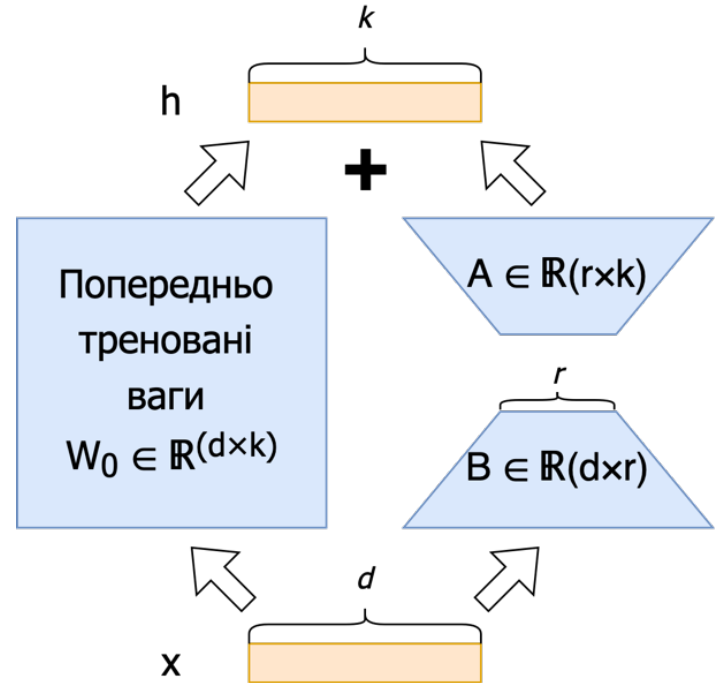
- Незважаючи на *стрімкий розвиток LLM*, їхня точність у вузькоспеціалізованих галузях може бути *обмеженою*.
- Одним із рішень цієї проблеми є *точне налаштування мовних моделей* під специфіку конкретної предметної області.
- Сучасні дослідження приділяють значну увагу методам точного налаштування, однак питання щодо його *практичної ефективності* для реальних, *практично значущих задач* з використанням обмежених ресурсів залишається *відкритим*.

Метод низькорангового адаптування (LoRA)

- $\Delta W \approx BA$

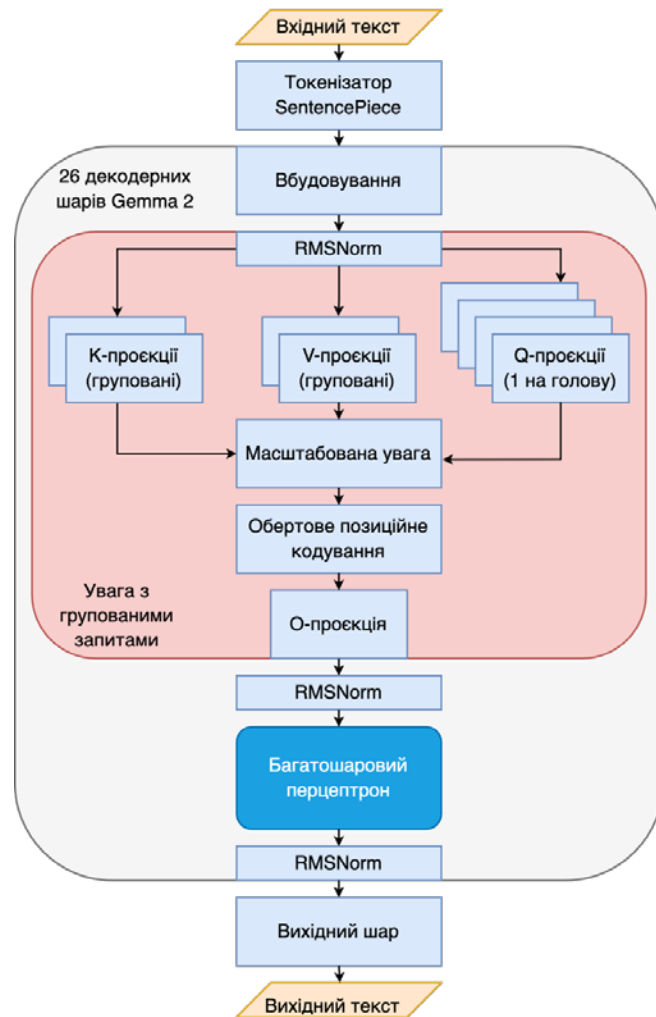
- $r \ll \min(d, k)$

- W_0 залишається замороженим під час навчання



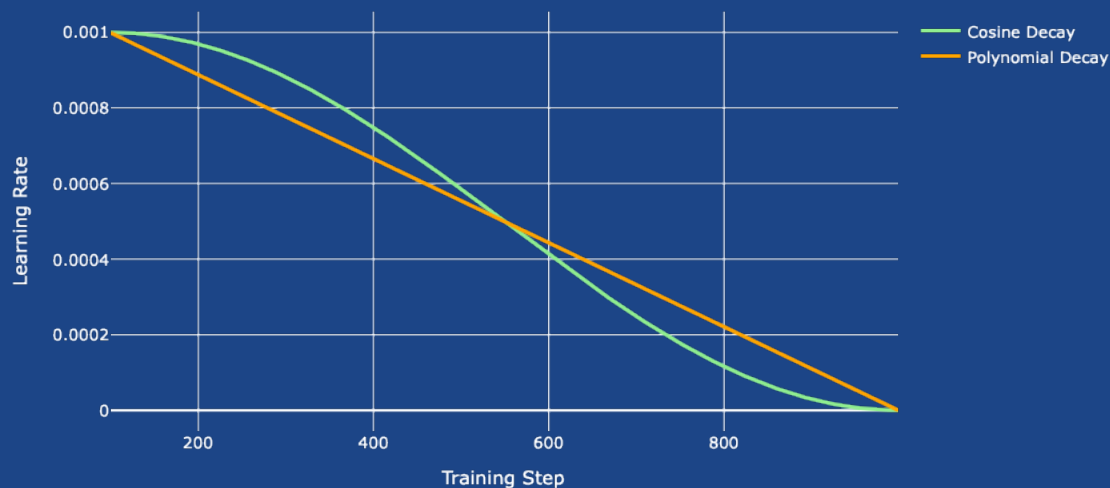
Архітектура Gemma 2 2B.

- 2.61 мільярди параметрів
- натренована на 2 трильйонах токенів
- контекстне вікно має довжину в 8192 токена



Графіки зниження темпу навчання для покращення точного налаштування

| Спад | Формула | Параметри |
|----------------|---|--|
| Поліноміальний | $lr = lr_{initial} * \left(1 - \frac{epoch}{max_epoch}\right)^{power}$ | lr – коефіцієнт швидкості навчання, $lr_{initial}$ – початковий коефіцієнт швидкості навчання |
| Косинусний | $\eta_t = \frac{\eta_0}{2} \left(1 + \cos\left(\pi C \frac{t}{T}\right)\right)$ | η_0 – початкова швидкість навчання, T – тривалість, C – масштаб. |



Опис предметних областей для точного налаштування



Науково-
теоретична
медицина

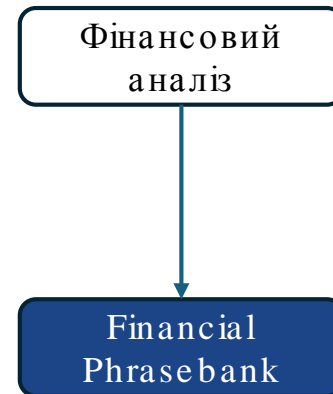
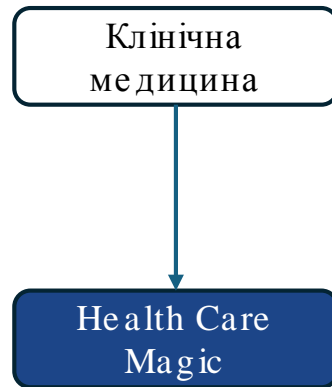
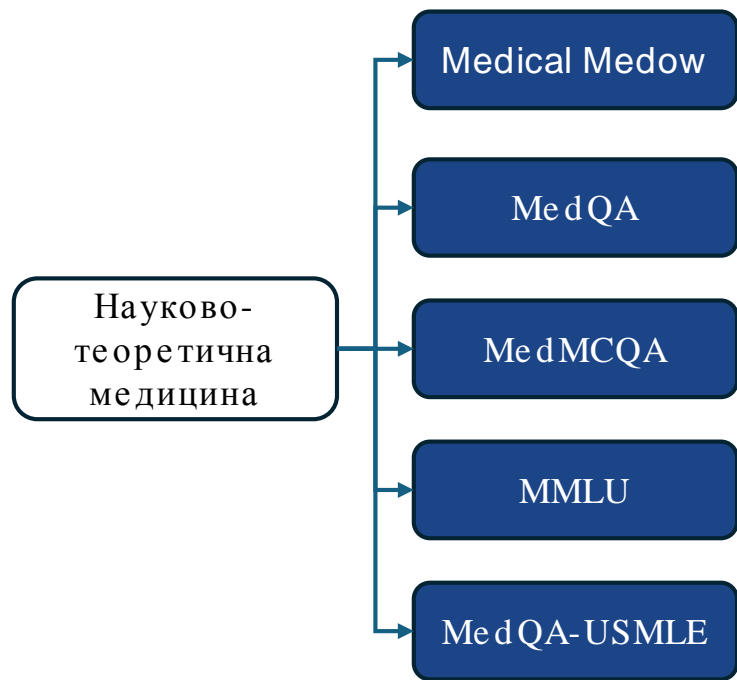


Клінічна медицина

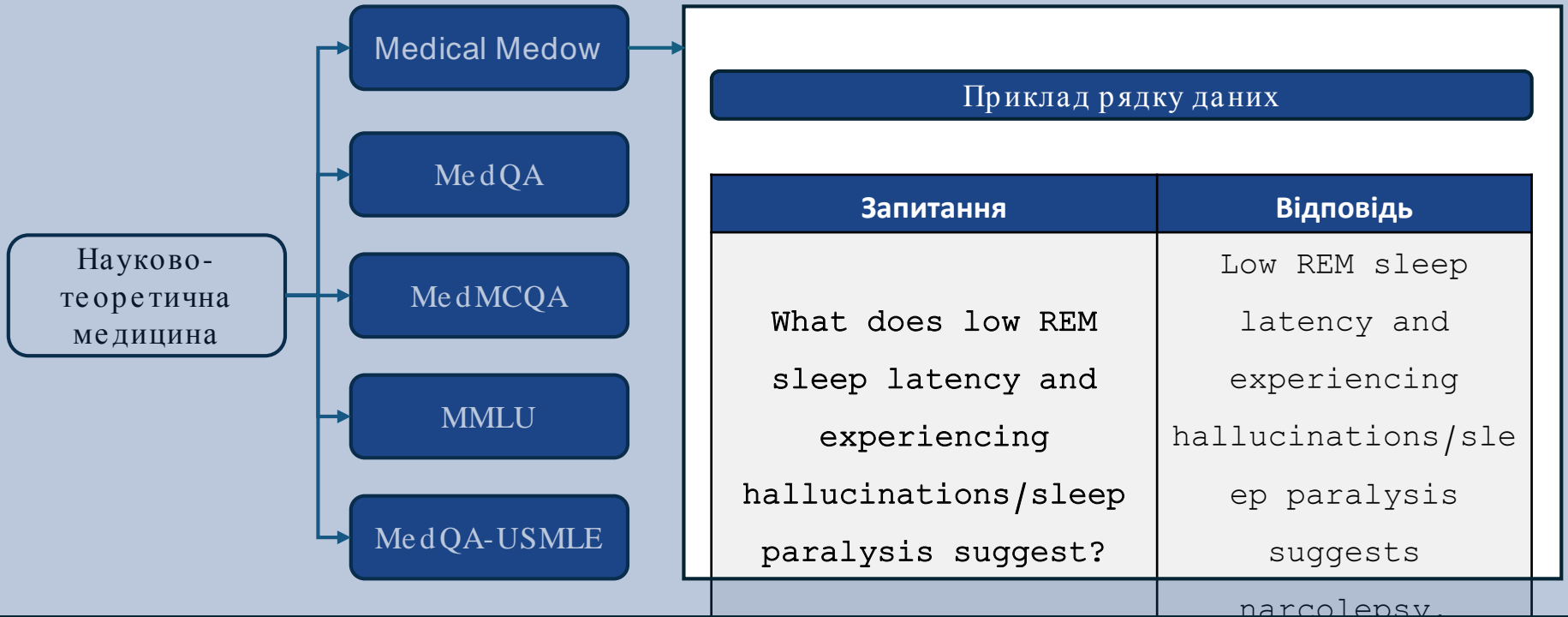


Фінансовий аналіз

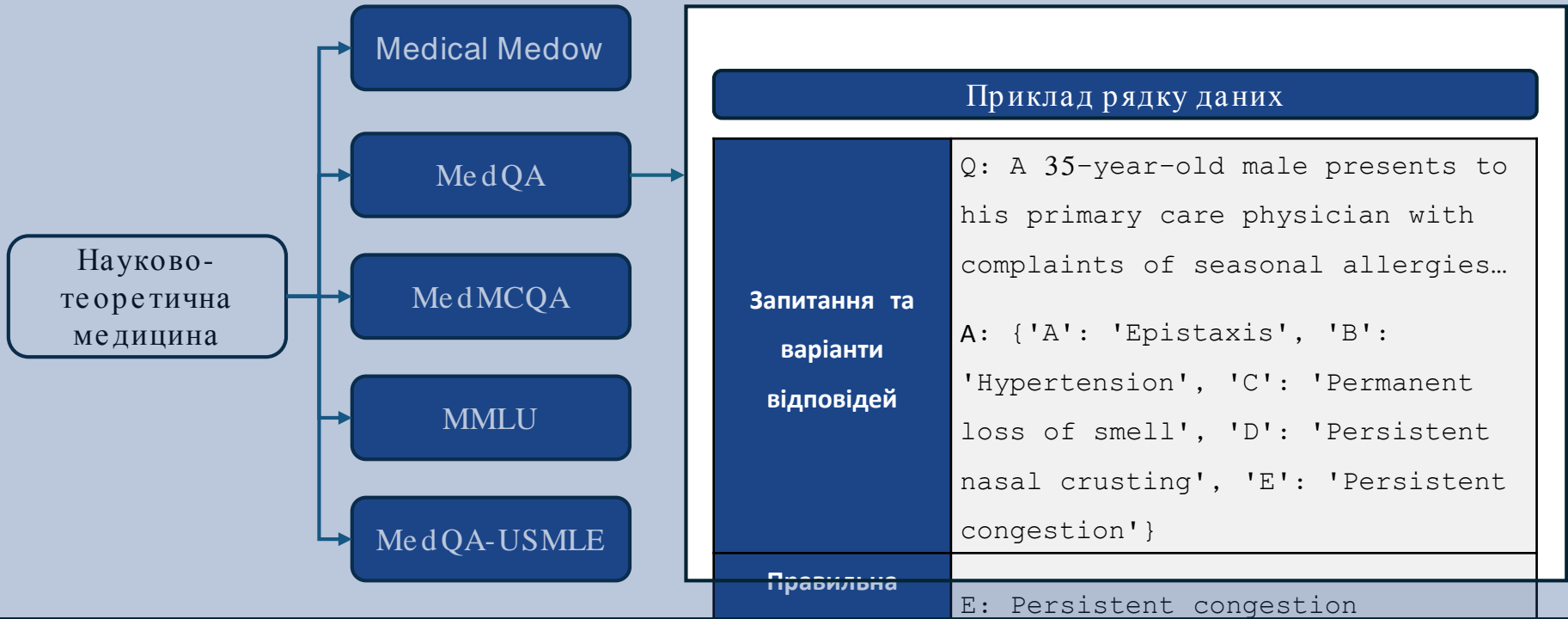
Набори даних



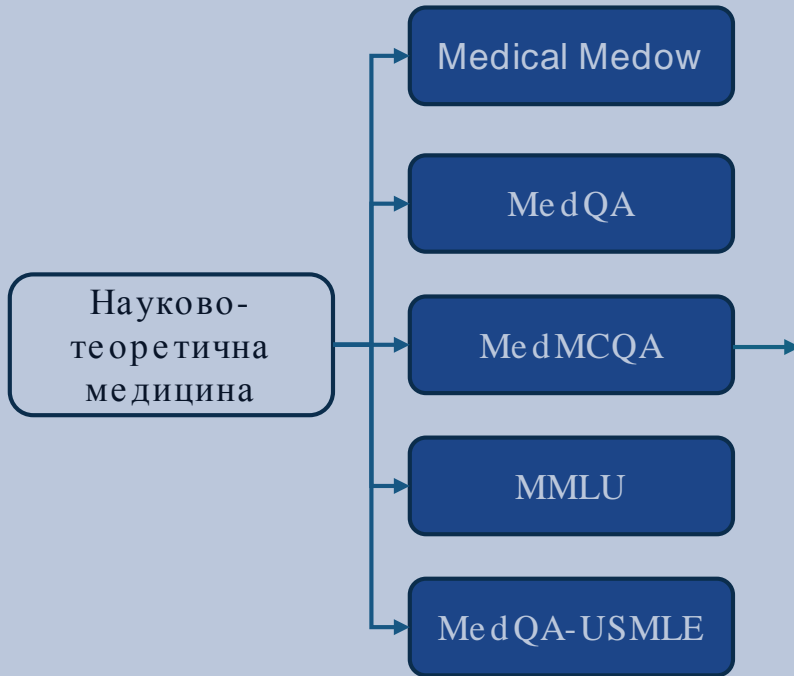
Набори даних



Набори даних



Набори даних



Приклад рядку даних

| Запитання | Варіант А | Варіант В | Варіант С | Варіант D | Відповідь |
|---|--------------------|------------------|-----------------|------------------|-----------|
| The pharmacokinetic change occurring in geriatric patient is due to ... | Gastric absorption | Liver metabolism | Renal clearance | Hypersensitivity | C |

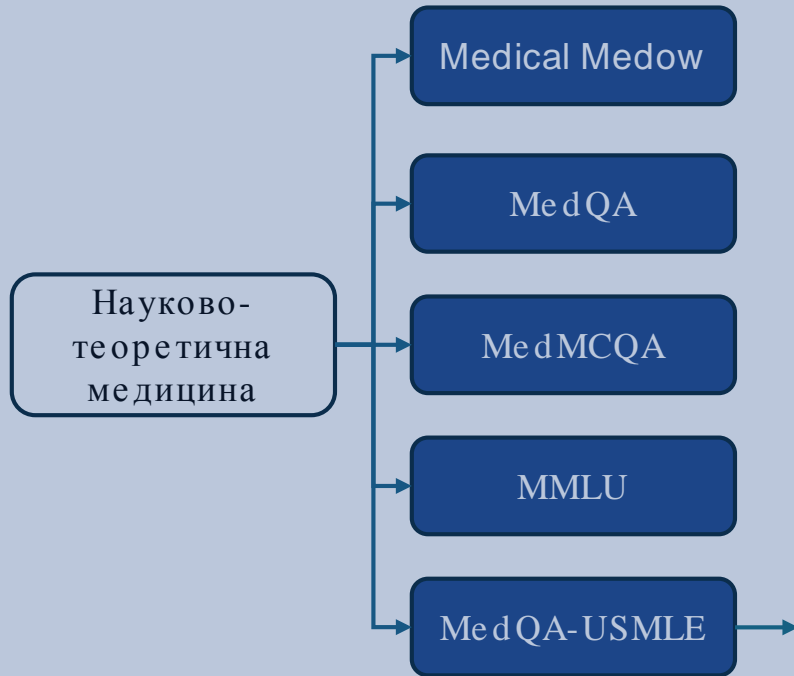
Набори даних



Приклад рядку даних

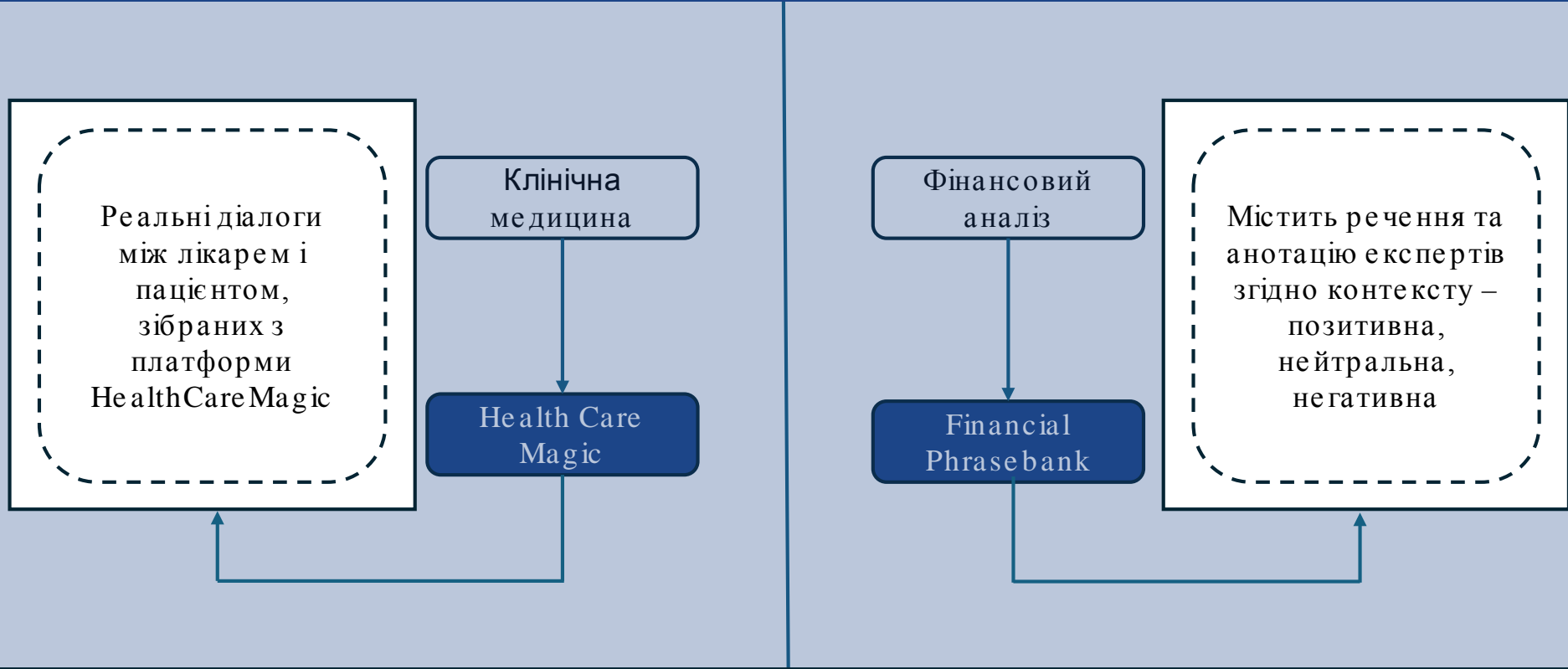
| Запитання | Область | Варіанти відповідей | Відповідь |
|---|------------------|--|-----------|
| Which of the following conditions does not show multifactorial inheritance? | Medical Genetics | {"Pyloric stenosis"; "Schizophrenia"; "Spina bifida (neural tube defects) "; "Marfan syndrome" } | D |

Набори даних



| Приклад рядку даних | | | | | |
|--|---|--|---------------------------------|---------------------------------------|-----------|
| Запитання | Варіант №1 | Варіант №2 | Варіант №3 | Варіант №4 | Відповідь |
| A 65-year-old male is treated for anal carcinoma with therapy... | Induces the formation of thymidine dimers | Induces the formation of disulfide bonds | Induces deamination of cytosine | Induces breaks in double-stranded DNA | 3 |

Набори даних



Точне налаштування для області науково-теоретичної медицини

| Етап | Тренувальні набори | Метод оптимізації | Коефіцієнт навчання | Ранг LoRA | Кількість епох | Розмір міні-пакету (batch size) | Проекційні шари з адаптерами | Тривалість навчання |
|------|--------------------------------|-------------------|--|-----------|----------------|---------------------------------|------------------------------|---------------------|
| 1 | Medical Meadow | AdamW (PyTorch) | З $1e-4$ до $1e-5$ з поліноміальним спадом | 512 | 1 | 8 | Всі | 39 хв. |
| 2 | MedMCQA, MedQA, Medical Meadow | AdamW (PyTorch) | Сталий $1e-5$ | 512 | 2 | 8 | Всі | 15 год. 9 хв. |

Застосування методу LoRA з рангом 512 дозволило зменшити кількість тренуваних параметрів на **74,58%** від загальної кількості.

Приклад використання MedGemmaScience

| Запитання | Відповідь |
|---|--|
| Question: Does the enzyme pyruvate carboxylase function in the glycolytic or gluconeogenic pathway, and what cofactor does it require? | Pyruvate carboxylase is an enzyme that functions in the gluconeogenic pathway, and it requires biotin as a cofactor. |
| Question: What is the first-line treatment for anaphylaxis? Options: A) Intravenous corticosteroids B) Oral antihistamines C) Intramuscular epinephrine D) Inhaled bronchodilators | C |

Точне налаштування для області клінічної медицини

| Тренувальні набори | Метод оптимізації | Коефіцієнт навчання | Ранг LoRA | Кількість епох | Розмір міні-пакету (batch size) | Проекційні шари з адаптерами | Тривалість навчання |
|--------------------|-------------------|---|-----------|----------------|---------------------------------|---|---------------------|
| Health Care Magic | AdamW (PyTorch) | Косинусний спад з початковим значенням $1e-5$ | 8 | 1 | 64 | Багатошаривий перцептрон, включно з розширенням та зменшенням розмірності | 2 год. 37 хв. |

Застосування методу LoRA з рангом 8 до вищезазначених шарів дозволило зменшити кількість тренуваних параметрів на **99,58%** від загальної кількості.

Точне налаштування для фінансової області

| Тренувальні набори | Метод оптимізації | Коефіцієнт навчання | Ранг LoRA | Кількість епох | Розмір міні-пакету (batch size) | Проекційні шари з адаптерами | Тривалість навчання |
|----------------------|-------------------|---------------------|-----------|----------------|---------------------------------|---|---------------------|
| Financial Phrasebank | AdamW (PyTorch) | Сталий 1e-5 | 512 | 2 | 16 | Всі, новий лінійний класифікаційний шар тренувався повністю | 2 хв 44 сек. |

Застосування методу LoRA з рангом 512 до вищезазначених шарів дозволило зменшити кількість тренуваних параметрів на **74,58%** від загальної кількості.

Результати тестування моделі MedGemmaScience

Результати тесту продуктивності MMLU для MedGemmaScience

| Область | Точність базової моделі, % | Точність після FT, % | Покращення Δ , % |
|--------------------------------|----------------------------|----------------------|-------------------------|
| Анатомія | 52,59 | 57,78 | +5,19 |
| Біологія рівня коледжу | 56,25 | 63,19 | +6,94 |
| Біологія рівня старшої школи | 64,19 | 68,39 | +4,2 |
| Вірусологія | 42,77 | 43,37 | +0,6 |
| Клінічні знання | 57,36 | 56,98 | -0,38 |
| Медична генетика | 61 | 65 | +4 |
| Медицина рівня коледжу | 53,18 | 55,49 | +2,31 |
| Професійна медицина | 42,28 | 51,47 | +9,19 |
| Професійна психологія | 48,37 | 50,16 | +1,79 |
| Психологія рівня старшої школи | 73,94 | 75,96 | +2,02 |
| Нутриціологія | 55,23 | 61,44 | +6,21 |
| Хімія рівня коледжу | 37 | 40 | +3 |
| Хімія рівня старшої школи | 35,96 | 39,9 | +3,94 |

Результати тестування на MedQA для MedGemmaScience

| Точність базової моделі, % | Точність після FT, % | Покращення Δ , % |
|----------------------------|----------------------|-------------------------|
| 32,91 | 42,66 | +9,75 |

Результати тестування моделі MedGemmaClinic

| Модель | BERTScore Precision | BERTScore Recall | BERTScore F1 |
|---------------------|---------------------|------------------|--------------|
| Базова Gemma 2 2B | 0,7670 | 0,8302 | 0,7972 |
| Gemma 2 2B після FT | 0.8614 | 0.8459 | 0.8535 |

Результати тестування моделі FinGemma

| Accuracy | Precision | Recall | F1-score |
|----------|-----------|--------|----------|
| 95,38% | 0,9339 | 0,9647 | 0,9485 |

Створення демонстраційних застосунків

Backend:

- FastAPI;
- Transformers та PyTorch;
- Uvicorn.

Frontend:

- HTML, TailwindCSS, FontAwesome;
- Javascript.

MedGemmaClinic Space

Головна сторінка MedGemmaClinic Space

MedGemmaClinic
AI Clinical Assistant

Disclaimer: MedGemmaClinic is not a substitute for professional medical advice, diagnosis, or treatment. Always consult a licensed healthcare provider before making health decisions. Do not ignore or delay medical care based on this tool.

Lately, I've been feeling constantly exhausted, even after a full nights rest, and I'm always thirsty. No matter how much water I drink, it never feels like enough. I've also been going to the washroom much more frequently, especially during the night, which is really disrupting my sleep. On top of that, I've noticed I'm losing weight without trying, my vision has become blurry at times, and I've had some tingling in my hands and feet. I'm starting to worry these symptoms might be connected. What do my symptoms suggest?

Hello! Thanks for your question on this chatbot. I can understand your concern. You have presented with multiple symptoms, including excessive thirst, frequent urination, blurred vision, fatigue, and weight loss. These are all suggestive of diabetes mellitus type 2. I would advise you to get a fasting blood sugar test done along with an HbA1c test. If both tests are positive, then you should start treatment immediately. The treatment includes lifestyle modifications (diet control, exercise), oral hypoglycemic agents, and insulin therapy if needed. Hope this helps. Take care.

Describe your symptoms or ask a medical question...

MedGemmaClinic | Created by Andrii Ostrolutskyi

MedGemmaClinic Space

Сторінка з параметрами генерації MedGemmaClinic Space

The screenshot shows the 'Advanced Settings' dialog for MedGemmaClinic. The interface has a blue header with the logo and 'AI Clinical Assistant' text. Below the header, there are four settings: 'Enable Sampling' (a toggle switch), 'Temperature' (a slider from 0.01 to 2.00), 'Top-p (nucleus)' (a slider from 0.01 to 1.00), and 'Repetition Penalty' (a slider from 1.00 to 2.00). At the bottom right, there are 'Cancel' and 'Apply Settings' buttons. The footer contains the text 'MedGemmaClinic | Created by Andrii Ostrolutskiy'.

MedGemmaClinic
AI Clinical Assistant

Advanced Settings

Enable Sampling
Toggle for deterministic or random responses

Temperature
Precise 0.60 Creative
Controls randomness (0 = deterministic, 2 = very random).

Top-p (nucleus)
0.01 0.60 1.00
Dynamic token selection based on cumulative probability.

Repetition Penalty
1.00 1.10 2.00
Penalize repeated tokens (higher = less repetition).

Cancel Apply Settings

MedGemmaClinic | Created by Andrii Ostrolutskiy

FinGemma Space

FinGemma
Financial News Sentiment Analyzer

Enter a news article URL to analyze financial sentiment

Article URL
 Analyze

Try: [BBC](#), [NY Times](#), or [CNN](#) financial news articles

Analysis Results

Overall Sentiment
Bearish (-0.46 sentiment score) 27%

| | | |
|--------------------------------|---------------------------------|-------------------------------|
| Positive Sentences 4 | Negative Sentences 15 | Neutral Sentences 6 |
|--------------------------------|---------------------------------|-------------------------------|

Article Analysis All Positive Negative Neutral

Asian shares slip as worries about U.S. debt send Wall St tumbling

ca.finance.yahoo.com - Original article

May 22, 2025

- TOKYO (AP) — Asian shares fell Thursday after Wall Street slumped under pressure from the Treasury bond market and worries about surging U.S. debt. -1.00 (100%)
- U.S. futures were little changed, while Japan's benchmark Nikkei 225 shed 1.0% in afternoon trading to 36,944.55. -0.84 (84%)
- Hong Kong's Hang Seng lost 0.9% to 23,615.21, while the Shanghai Composite edged down 0.1% to 3,383.10. -1.00 (100%)
- Australia's S&P/ASX 200 slipped 0.5% to 8,342.80. -1.00 (100%)
- South Korea's Kospi dropped 1.1% to 2,595.69. -1.00 (100%)
- Rising yields for U.S. Treasury bonds are a canary in the coal mine, Stephen Innes of SPI Asset Management said in a commentary. -0.67 (78%)
- The U.S. still has the biggest markets, the deepest liquidity, and the dollar's inertia working in its favor. 0.81 (81%)
- But even inertia can't outrun compound interest and structural deficits forever," he wrote. -0.47 (53%)

Висновки

В результаті виконання кваліфікаційної роботи було:

- отримано три моделі, які демонструють *кращі результати* тестування ніж базові, що доводить доцільність точного налаштування *в умовах обмежених обчислювальних ресурсів*;
- досягнуто збільшення точності на **9,75%** для науково-теоретичної медицини, влучності на **9.44%** для клінічної медицини, та точність **95,38%** для фінансової моделі.
- доведено ефективність використання високого рангу (**512**) для якісної передачі предметних знань;
- запропоновано *двоет апне низькорангове адаптування* для медичної сфери;
- показано ефективність використання низькорангових адаптерних матриць виключно для шарів багат шарового перцептрону;
- розроблено демонстраційні застосунки для клінічної та фінансової сфер.

Дякую за увагу!

