



# Кваліфікаційна робота бакалавра

Розробка застосунку для генерації SQL-запитів,  
заданих природною мовою

Виконав:

ст. гр. ІТШІ-21-5

Кальченко А. С.

Керівник:

доц. Вітько О. В.

# Актуальність дослідження

## Проблематика:

- Залучення висококваліфікованих експертів до вирішення рутинних завдань призводить до нераціонального використання ресурсів

## Обмеження традиційного підходу:

- Необхідність знання SQL та структури БД
- Високий поріг входу для нетехнічних користувачів

## Рішення:

- Генеративна нейромережа для перетворення природної мови в SQL

## Очікуваний ефект:

- Розширення кола користувачів
- Інтуїтивний, зручний інтерфейс





# Наявні аналоги

## 1. API / Моделі

*Для вбудовування в застосунки або бекенд*

- OpenAI GPT-4 / Codex – точна генерація SQL, API-доступ
- SQLCoder (Defog) – open-source, оптимізований для NL2SQL
- SmBoP / PICARD (Salesforce) – SOTA-архітектури

## 2. Браузерні застосунки

*Готові інтерфейси без потреби у коді*

- Text2SQL.AI – простий інтерфейс, підтримка СУБД
- AI2SQL – SaaS з поясненням запитів
- SQL Chat – open-source застосунок із GPT-інтеграцією

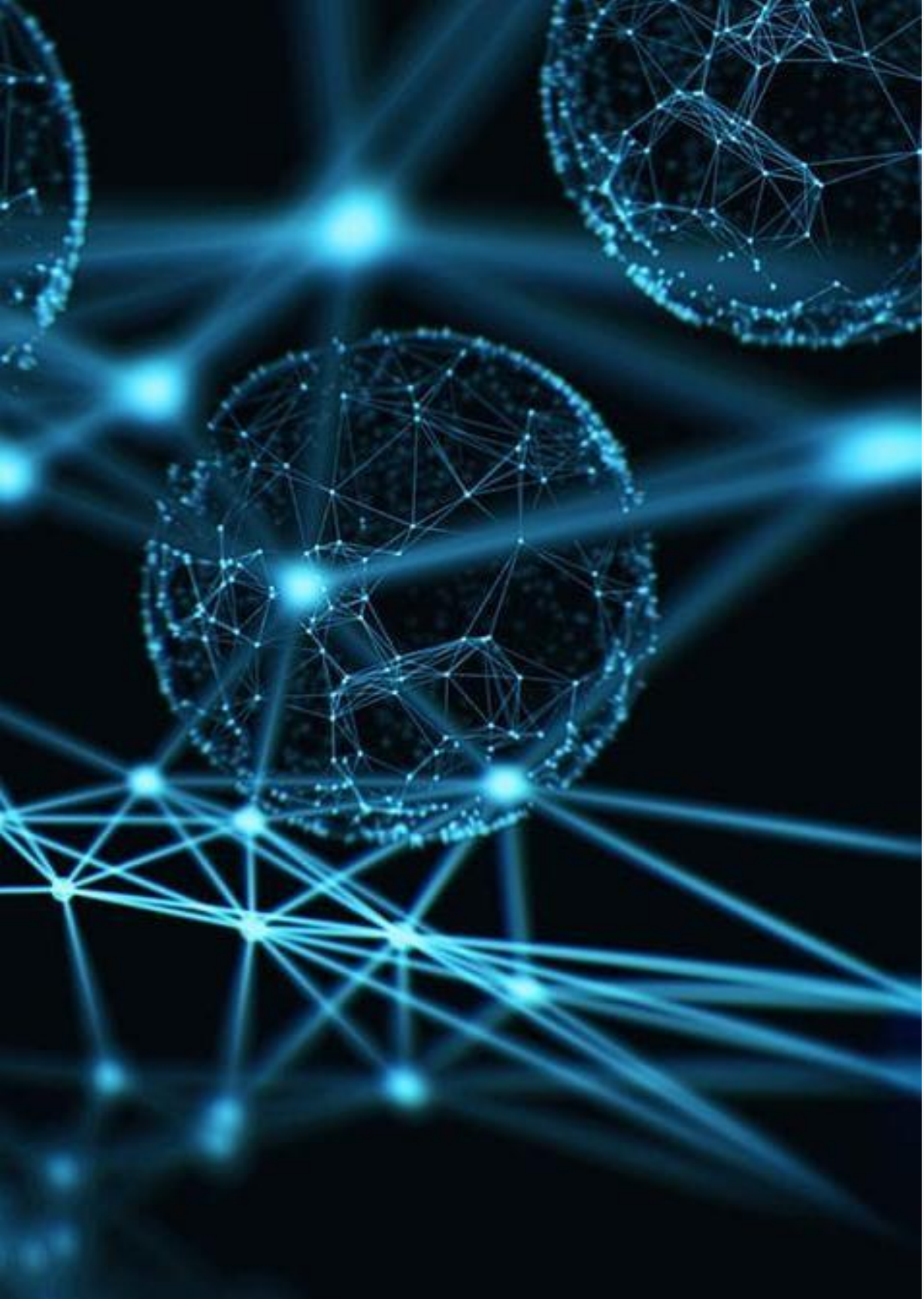
## 3. Вбудовані в СУБД / BI-системи

*Інтеграція ШІ в звичне середовище роботи з даними*

- Power BI (Q&A) – запити природною мовою у звітах
- Google BigQuery + Vertex AI – кастомна NL2SQL-платформа
- ThoughtSpot – бізнес-аналітика з природномовними запитами







# Дослідження моделей для вирішення поставленої задачі

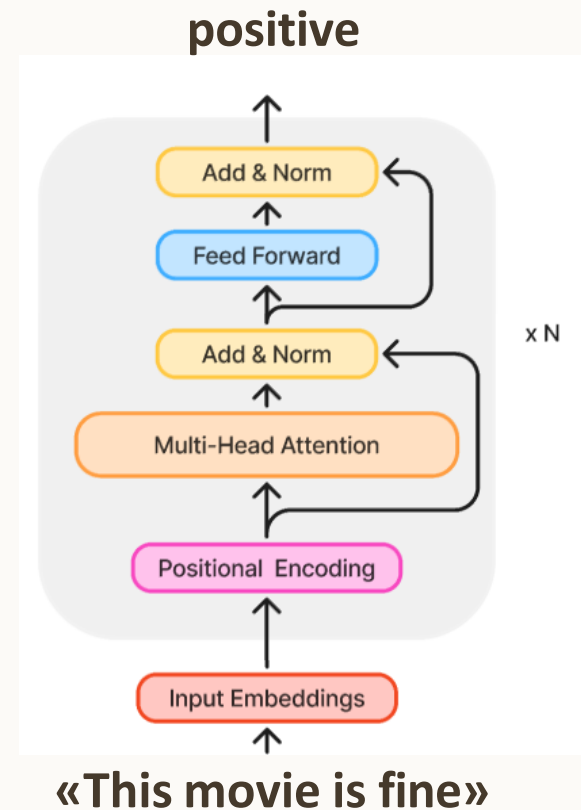
## Обмеження RNN / LSTM:

- Послідовна обробка токенів: унеможливорює паралельне навчання → повільно навчаються
- Проблеми з довгими залежностями: труднощі в урахуванні інформації з початку запиту
- Обмежений контекст: модель «забуває» далекі зв'язки в запиті
- Гірші результати в складних задачах NLP, зокрема у синтаксично точних запитах як SQL

## Переваги трансформерів:

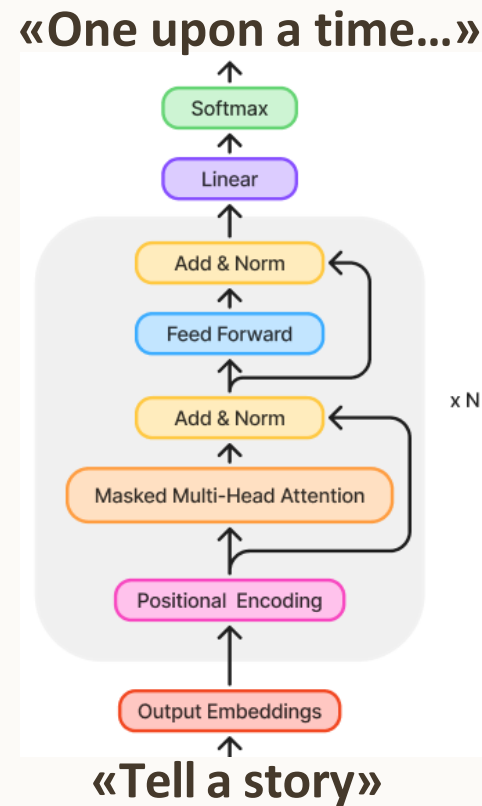
- Паралельне навчання: завдяки self-attention — швидше на порядок
- Глобальний контекст: модель бачить увесь запит одночасно, що важливо для SQL-структур
- Краще моделювання складної структури запиту
- SOTA результати: сучасні моделі NL2SQL базуються саме на трансформерах

# Аналіз трансформер-моделей



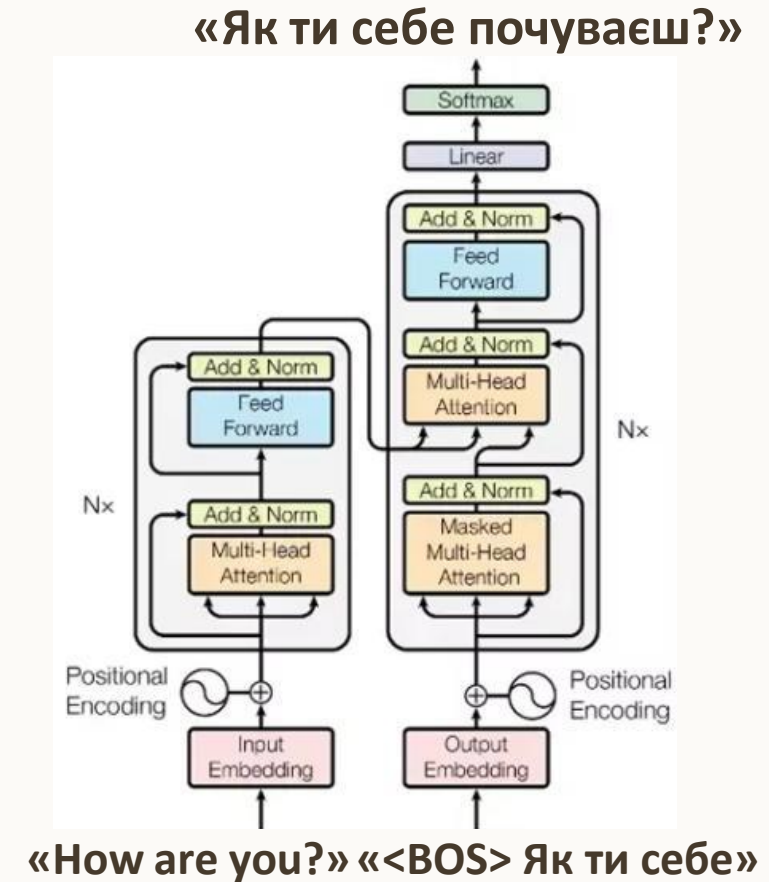
*Encoder-only*

Фокусується на розумінні вхідних даних.



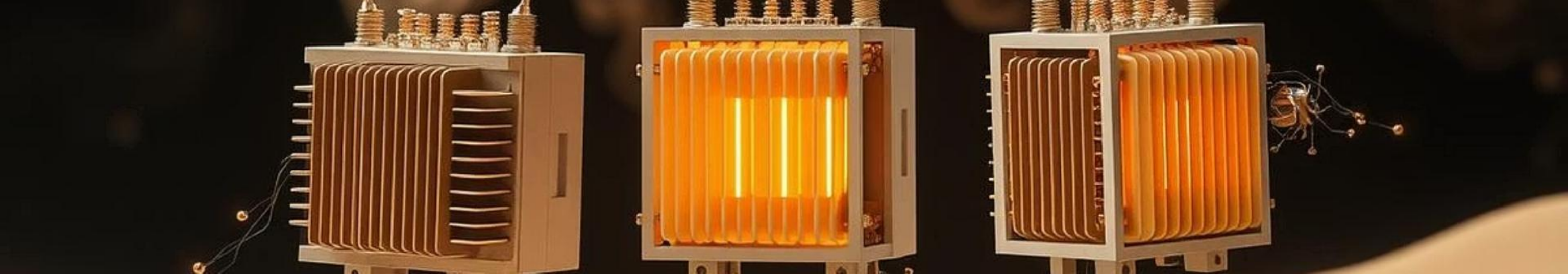
*Decoder-only*

Фокусується на генерації.



*Encoder-decoder*

Комбінує розуміння та генерацію.



# Вибір моделі для генерації запитів

## T5

Універсальна модель для перекладу, узагальнення та відповідей на запитання.

## BART

Ефективна для узагальнення та генерації зв'язного тексту.

## Pegasus

Спеціалізується на узагальненні довгих текстів.

## Marian MT

Використовується для точного перекладу природних мов.

# Вибір метрики оцінки

## ROUGE-S ✗

- + Гнучка у завданнях перефразувань
- Не враховує порядок слів

## ROUGE-N ?

- + Швидка в обробці
- Негнучка на важких послідовностях

## ROUGE-L ?

- + Враховує порядок токенів
- Не враховує малі підпослідовності

## BLEU ✓

- + Штрафує за надто короткі відповіді
- Доволі сувора на малих послідовностях

## METEOR ✗

- + Враховує спектр варіацій слова та порядок слів
- В задачах із суворим синтаксисом варіативність заважає



# Аналіз метрики BLEU

$\text{BLEU} = \text{BP} \cdot \exp \left( \sum_{n=1}^N w_n \log(p_n) \right)$	$\text{BP} = \begin{cases} 1, & \text{якщо } c > r \\ \exp \left( 1 - \frac{r}{c} \right) & \text{— інакше} \end{cases}$	$\text{Precision} = \frac{C_R}{C}$
--	--	------------------------------------

$p_n$  – precision для n-грам (зазвичай 4)

$w_n$  – вектор ваг для n-грам (зазвичай  $\frac{1}{N}$ );

BP – Bravery Penalty, штраф за короткі відповіді;

$r$  – довжина еталонного речення;

$c$  – довжина згенерованого речення.

$C_R$  – кількість n-грам, зі згенерованої послідовності, які співпадають з n-грамами із еталонної;

$C$  – кількість всіх n-грам в згенерованій послідовності;



# Аналіз обраного набору даних

Колонка	Кількість унікальних	Опис
domain	100	Тематика запиту (здоров'я, фінанси, космос)
sql_complexity	8	Тип запиту
sql_task_type	4	Тип задачі запиту
sql_prompt	100000	Запит англійською
sql_context	89766	Схема БД
sql	99271	Запит SQL
domain_description	100	Опис відповідної колонки датасету
sql_complexity_description	8	
sql_task_type_description	4	
sql_explanation	99777	

## Приклади входжень колонок:

### sql\_context

- CREATE TABLE Employees (id INT, first\_name TEXT, last\_name TEXT, salary FLOAT)

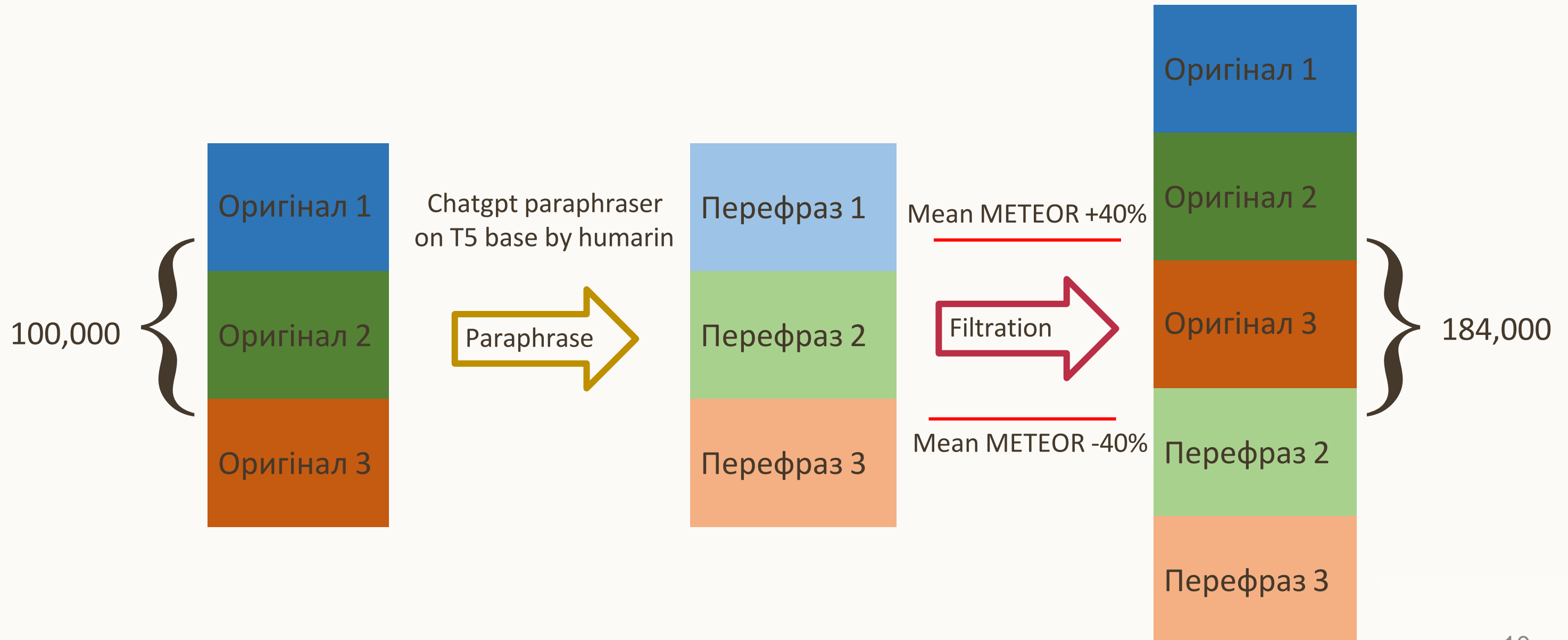
### sql\_prompt

- List all employees with a salary more than 19000

### sql

- SELECT first\_name, last\_name, salary FROM Employees WHERE salary > 19000

# Перефразування та фільтрація за допомогою METEOR



# Перевірка гіпотези про перефразування

Model	Score type	Dataset type	Score
ParaBART	Mean	train	0,684
		test	0,686
	Median	train	0,701
		test	0,703
RawBART	Mean	train	0,660
		test	0,664
	Median	train	0,666
		test	0,672

Висновки:

- Моделі не перенавчились (оцінка на тренувальному наборі і тестовому не дуже різниться)
- Модель із перефразуваннями навчилась краще за модель без них



# Оптимізація гіперпараметрів

## learning\_rate



- Швидкість оновлення ваг
- Типовий діапазон значень –  $[1e-5, 5e-4]$
- Оптимальне значення для T5 - **0.005**
- Оптимальне значення BART - **3.73**

## number\_of\_epoch



- Кількість проходів по всіх даних
- Типовий діапазон значень –  $[1, 20]$
- Оптимальне значення для T5 – **2**
- Оптимальне значення BART - **2**

## batch\_size



- Кількість прикладів за одну ітерацію
- Типовий діапазон значень –  $[8, 256]$
- Оптимальне значення для T5 – **32**
- Оптимальне значення BART - **64**

## warmup\_ratio



- Частка кроків «повільного старту»
- Типовий діапазон значень –  $[0.01, 0.5]$
- Оптимальне значення для T5 - **0.19**
- Оптимальне значення BART - **0.25**

## weight\_decay

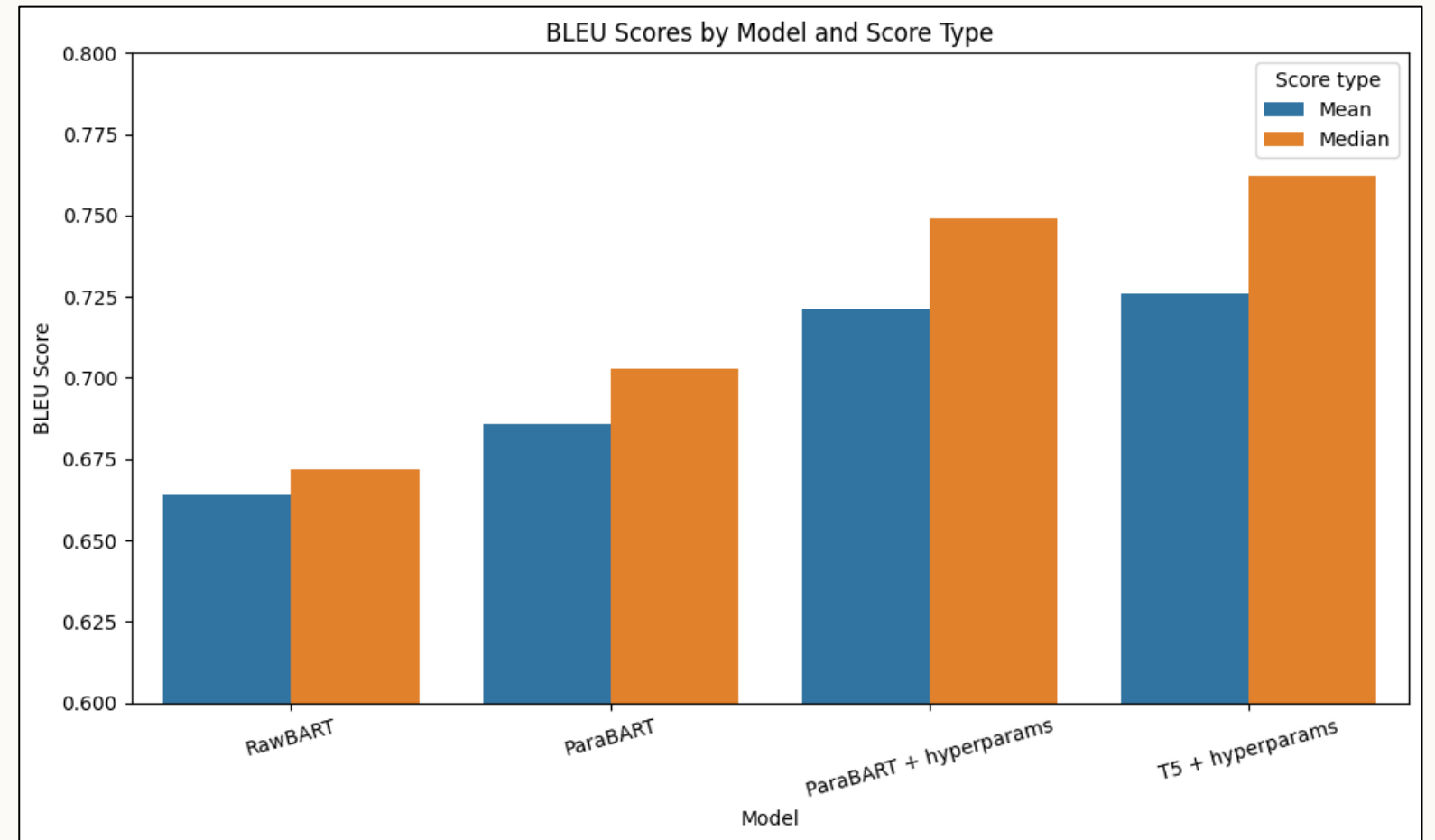


- L2 регуляризація
- Типовий діапазон значень –  $[0, 0.4]$
- Оптимальне значення для T5 - **0.067**
- Оптимальне значення BART – **0.24**

# Фінальне порівняння моделей

## Висновки:

- Найкращою моделлю є T5 із використанням датасету із перефразуваннями та оптимізацією гіперпараметрів
- Через велику різницю між медіанним та середнім значенням є висока ймовірність викидів серед дуже низьких значень

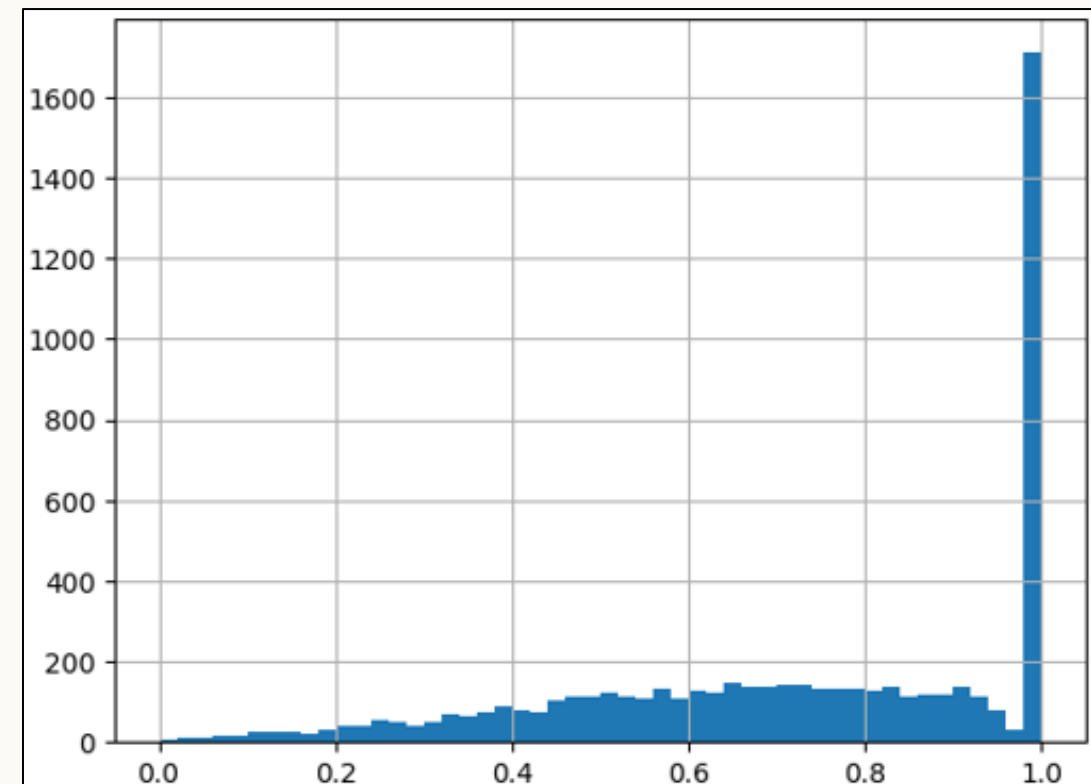


# Аналіз результатів оцінки T5

Відсоток прикладів із оцінкою 1  
за класами колонки sql\_complexity

sql_complexity	
basic SQL	0.474381
aggregation	0.203221
single join	0.065463
subqueries	0.054404
set operations	0.039216
multiple_joins	0.014388
CTEs	NaN
window functions	NaN

Name: count, dtype: float64





# Перевірка гіпотези щодо простоти класів

Відсоток прикладів із оцінкою 1 за класами колонки sql\_complexity

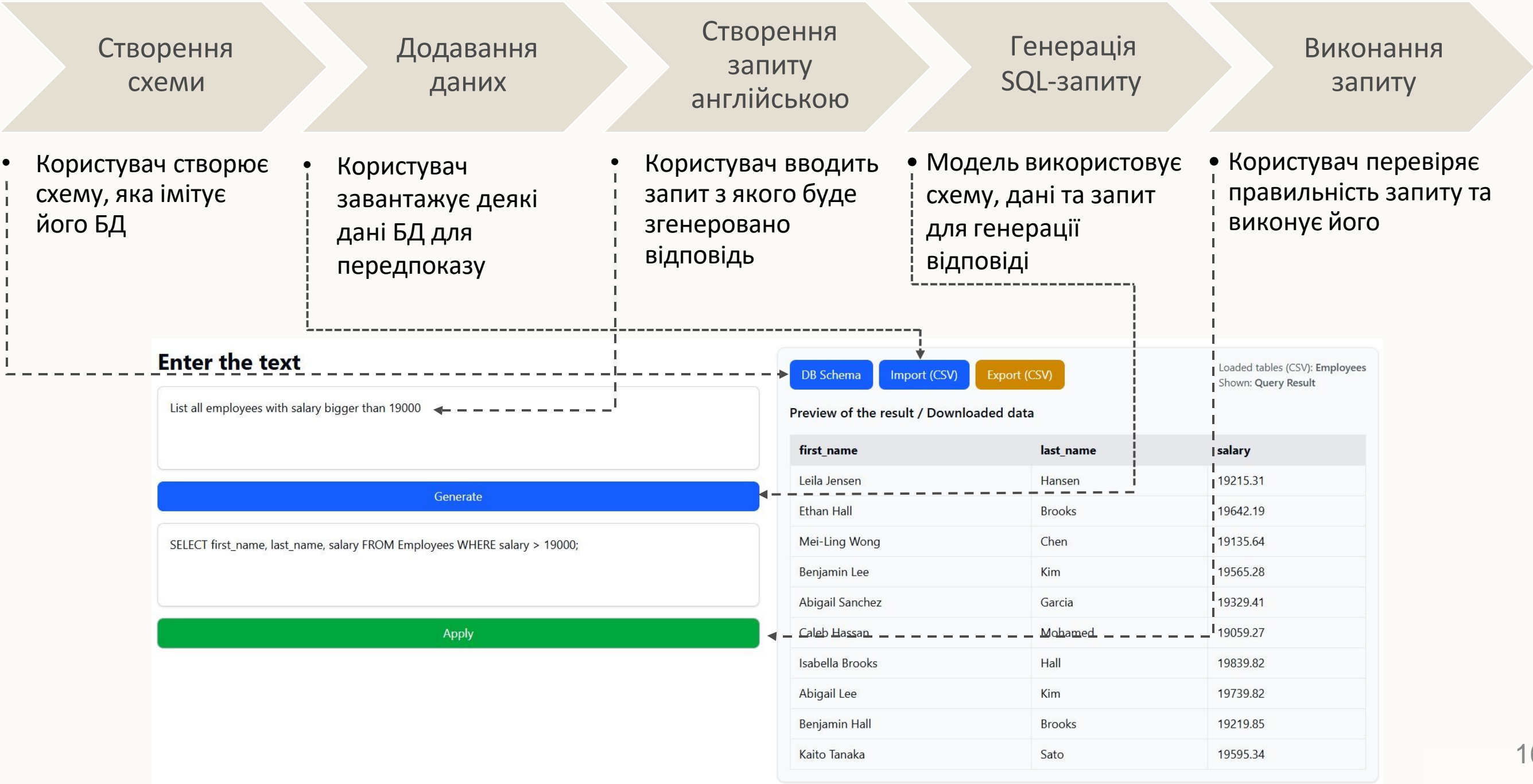
basic SQL	0.474381
aggregation	0.203221
single join	0.065463
subqueries	0.054404
set operations	0.039216
multiple_joins	0.014388
CTEs	NaN
window functions	NaN

- Кореляційний аналіз показав значення **-0.87**

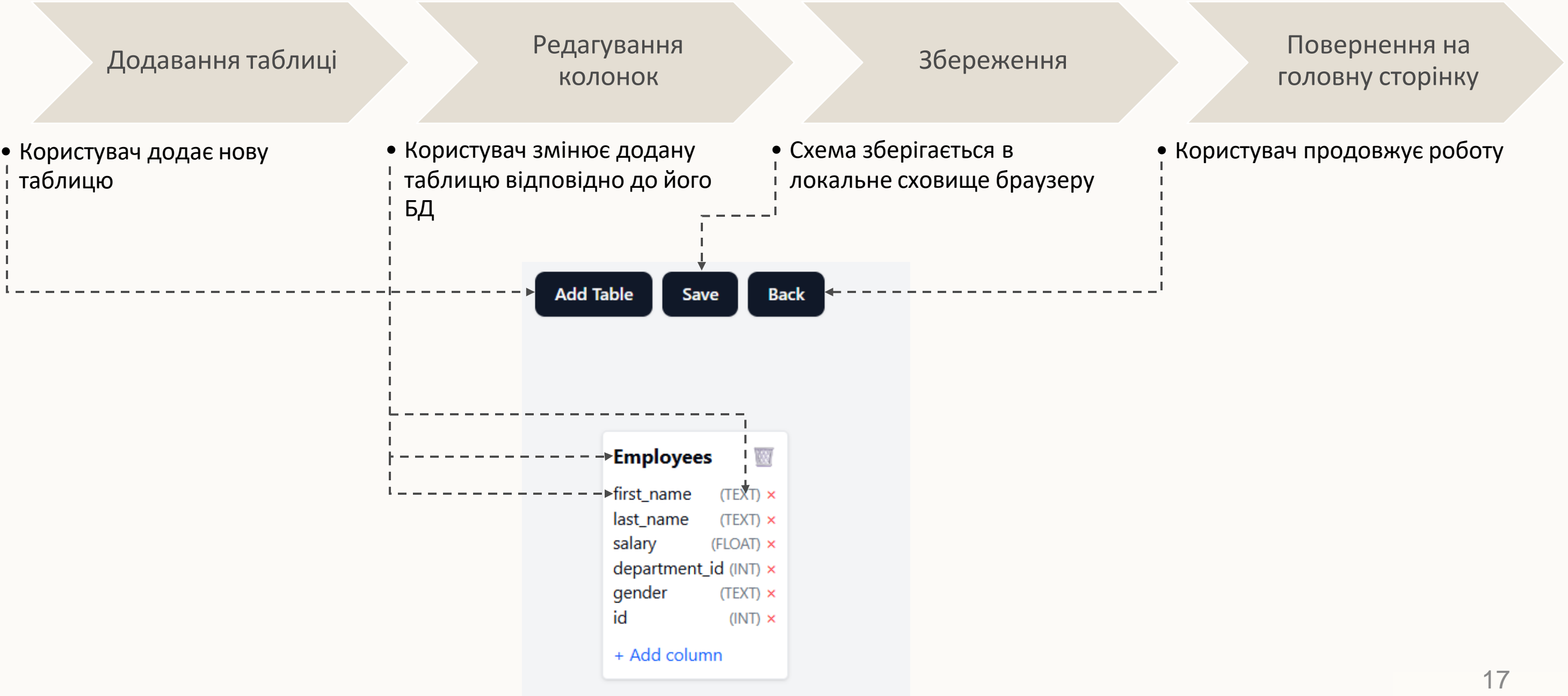
Середня кількість слів у класах колонки sql\_complexity

basic SQL	11.275706
aggregation	15.103528
single join	20.931151
CTEs	21.000000
subqueries	24.077720
window functions	24.668367
set operations	26.431373
multiple_joins	32.417266

# Етапи роботи користувача із головною сторінкою застосунку



# Етапи роботи із сторінкою редагування схеми БД





# Використані інструменти

Датасет

The logo for Gretel, featuring the word "gretel" in a bold, blue, lowercase sans-serif font.

Обробка даних



Клієнтська частина

The logo for Next.js, featuring the word "NEXT" in a large, black, uppercase sans-serif font, with ".js" in a smaller font size to the right. A diagonal line crosses through the "X".

Моделі для навчання



Серверна частина



Оптимізація гіперпараметрів



# Висновки

Результатом роботи є браузерний застосунок із клієнтською та серверною частиною, що може перетворювати запити природною мовою на SQL.

## Переваги

- В застосунку використовується модель власного донавчання.
- Модель для генерації знаходиться в репозиторії HuggingFace (ArtemKalchenko/t5-small\_for\_sql\_generation).

## Було проаналізовано

- Метрики: BLEU, ROUGE, METEOR
- Моделі: Pegasus, T5, BART, Marian MT

## Серед моделей обрано і протестовано:

- T5: медіанна точність за метрикою BLEU-2 – **0.762**
- BART: медіанна точність за метрикою BLEU-2 – **0.749**

## Перспективи

- Доновчання моделі для роботи із складнішими запитами.
- Інтеграція в СУБД.
- Навчання моделі задля роботи із запитами іншими мовами.
- Додавання підтримки бази даних.

 **Обрано модель T5 для генерації SQL-запитів у застосунку**