

СЕКЦІЯ «ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ СИСТЕМИ»

Христина Бабій, Юрій Процик*

¹ НЛТУ України, Львів, Україна, kh.babiy@nltu.lviv.ua

² НЛТУ України, Львів, Україна, protsyk@nltu.edu.ua

Використання методів NLP для розширеного пошуку та аналізу текстових даних у спортивних інформаційних системах

***Анотація.** Досліджено ефективність великих мовних моделей (LLMs) для задачі Text-to-SQL на прикладі україномовних запитів у спортивній інформаційній системі. Проведено порівняльний аналіз шести моделей із використанням стратегії Few-shot Prompting за точністю, повнотою, часом виконання та обчислювальною вартістю генерації SQL-запитів. Результати показали, що GPT-4.1 досягає найвищої повноти (0.95), тоді як Gemini 2.0 Flash є найбільш економічним рішенням для практичного застосування.*

***Ключові слова** – Text-to-SQL, велика мовна модель, Prompt Engineering, Few-shot Prompting, спортивна інформаційна система.*

Вступ. Стрімке зростання обсягів даних перетворило інформацію на стратегічний ресурс у всіх сферах людської діяльності. У спеціалізованих галузях, таких як спортивні інформаційні системи, масивні бази даних містять важливі відомості, доступ до яких вимагає оперативності та високої точності. Однак, щоб взаємодіяти з цими даними, користувачам необхідно володіти мовою структурованих запитів (SQL), що ускладнює доступ до інформації для звичайних користувачів та створює додаткове навантаження для фахівців, яким доводиться створювати численні запити.

Проблема інтеграції природної мови зі структурованими даними зумовила появу задачі Text-to-SQL [1]. Text-to-SQL – це фундаментальна задача, яка перетворює запити, сформульовані природною мовою (Natural Language Queries), на команди мови структурованих запитів.

Активний розвиток LLMs спричинив зміни в архітектурі та принципах побудови систем Text-to-SQL, визначивши новий етап у дослідженнях цієї га-

лузі. Завдяки здатності до глибокого розуміння природної мови, LLMs забезпечують ефективну генерацію SQL-запитів. Для підвищення точності застосовуються спеціалізовані стратегії [2], такі як Prompt Engineering (наприклад, Zero-shot, Few-shot або Chain of Thought) та Fine-tuning (повне або параметрично-ефективне). Ці методи допомагають вирішувати складні проблеми Text-to-SQL, включаючи неоднозначність (амбівалентність) природної мови, правильне підключення до таблиць та обробку складних SQL-операцій (JOIN, вкладені підзапити, тощо).

Розробка інтелектуальної системи. Метою роботи є розробка інтелектуальної системи на основі NLP та LLMs для інтерпретації користувацьких запитів, автоматичного формування SQL-запитів або виконання вебпошуку та надання структурованих та зрозумілих відповідей. Для забезпечення високої точності Text-to-SQL перетворення та класифікації джерела інформації (структурована база даних чи загальний вебпошук) було обрано стратегію Few-shot Prompting, без виконання Fine-tuning: модель залишається незмінною (Zero-shot), але отримує приклади взаємодії у запиті. Такий підхід дозволяє моделям краще інтерпретувати складні та вузькоспеціалізовані запити, а також правильно відтворювати бажаний формат виведення.

Промпт моделі був структурований для забезпечення високої якості генерації, а саме, включав:

- Чітке визначення завдання: аналіз запиту, класифікація на “Database” або “Internet” та генерація відповіді у форматі JSON.
- Надання повної схеми таблиць з первинними/зовнішніми ключами та додатковими відомостями про домен фігурного катання.
- Включення 6-ти конкретних прикладів взаємодії (запит користувача – очікуваний JSON-об’єкт). Ці приклади демонстрували бажаний формат виведення, логіку складних SQL-операцій (JOIN, ORDER BY, LIMIT) та необхідність класифікації джерела інформації.

Експериментальна частина. Для оцінки точності та надійності функціонування інтелектуальної пошукової системи на базі мовних моделей, було

проведено серію з 30 експериментів. До порівняння залучено шість потужних мовних моделей від різних провайдерів: Gemini 2.5 Flash Preview, Gemini 2.0, GPT-4.1, GPT-4.1-nano, Claude Opus 4, Claude 3.5 Haiku. Кожна з моделей отримала однакову групу із 5 запитань українською мовою, що імітують запити кінцевого користувача у сфері фігурного катання. Підібрані запити охоплюють різні типи операцій над базою даних, зокрема фільтрацію, агрегацію, сортування та обчислення максимуму/мінімуму.

Якість відповіді вимірювалася двома метриками: точність (визначалася вручну на основі трьох рівнів відповідності (100%, 50-99%, <50%), де 100% означає повну синтаксичну та логічну коректність SQL-запиту, що призводить до релевантного результату) та повнота (відображає, наскільки згенерований запит охоплює всі необхідні умови та фільтри, сформульовані у запитах). Час генерації (сек) фіксує швидкість виконання завдання Text-to-SQL. Вартість (\$) показує фактичні витрати на токени, що були використані для обробки запиту та генерації відповіді. Нижче наведена таблиця усереднених результатів дослідження.

Таблиця 1 – Таблиця результатів аналізу

Назва моделі	Середня точність	Середня повнота	Середній час генерації (сек)	Середня кількість токенів	Середня вартість (\$)
Claude-opus-4	1	0.85	4.568	2636.4	0.045126
Claude-3-5-haiku	0.9	0.85	2.186	2642.4	0.0024308
Gemini 2.5 Flash Prev	1	0.7	7.28	1794.2	0.001365
Gemini 2.0 Flash	1	0.85	2.178	1779	0.000935
GPT-4.1	1	0.95	1.728	1788.8	0.004084
GPT-4.1-nano	0.4	0.22	1.838	1772	0.0001974

Висновок. Використання методів NLP для розширеного пошуку та аналізу текстових даних у спортивних інформаційних системах дає змогу спростити

та прискорити доступ користувачів до важливих даних, автоматизуючи процес перетворення запитів природною мовою (Text-to-SQL) у точні команди SQL. Завдяки впровадженню великих мовних моделей у поєднанні зі стратегією Few-shot Prompt, забезпечується висока точність обробки запитів. На основі цього аналізу, GPT 4.1 та Gemini 2.0 Flash виділяються як найбільш ефективні моделі для генерації SQL-запитів на основі запитів користувачів у сфері фігурного катання. GPT-4.1 демонструє найвищу загальну якість відповідей завдяки високому показнику повноти (0.95) та загальної точності, що робить її найкращим вибором у сценаріях, де пріоритетом є достовірність і повнота результатів. Gemini 2.0 Flash пропонує дуже близьку якість при значно нижчій ціні та подібній швидкості, що робить її найбільш привабливим варіантом для практичного застосування в інтелектуальних пошукових системах, де співвідношення ціна/якість/швидкість є вирішальним фактором.

Список використаних літературних джерел

- [1] Zhu, X., Li, Q., Cui, L., & Liu, Y. (2024). Large Language Model Enhanced Text-to-SQL Generation: A Survey. arXiv preprint arXiv:2410.06011. <https://arxiv.org/abs/2410.06011>
- [2] Iusztin, P., & Labonne, M. (2024). The LLM Engineering Handbook. Packt Publishing.

Using NLP methods for advanced search and analysis of textual data in sports information systems

Khrystyna Babii, Protsyk Yurii

The effectiveness of Large Language Models (LLMs) for the Text-to-SQL task was investigated using the example of Ukrainian-language queries in a sports information system. A comparative analysis of six models was conducted using the Few-shot Prompting strategy, evaluating accuracy, recall, execution time, and computational cost of SQL query generation. The results showed that GPT-4.1 achieves the highest recall (0.95), while Gemini 2.0 Flash is the most economical solution for practical application.