Цап Владислав Богданович, аспірант
Національний університет «Львівська політехніка», Львів
<https://orcid.org/0000-0002-8062-0079>

**МОЖЛИВОСТІ ІНТЕГРАЦІЇ ГРАФОВИХ БАЗ ДАНИХ З ВЕЛИКИМИ МОВНИМИ МОДЕЛЯМИ**

Останній стрибок у розвитку великих мовних моделей (ВММ) супроводжується їхнім все більш широким впровадженням у різноманітні сфери діяльності, що забезпечує зручні інтерфейси для користувачів. Проте, існують значні обмеження, які потребують подальшого дослідження та вирішення [1]. Зокрема, для застосування ВММ у сферах, що вимагають високої точності, критично важливим є забезпечення можливості відстеження та інтерпретації процесу прийняття рішень, що лежить в основі генерації відповіді. Це дозволить зрозуміти обґрунтування та логіку, що призвели до конкретного результату. Крім того, надзвичайно велика кількість параметрів, характерна для сучасних ВММ, зумовлює значні обчислювальні вимоги та потребує значних ресурсів. Тому, оптимізація архітектури та алгоритмів, спрямована на зменшення обчислювального навантаження, є актуальним завданням.

Проаналізувавши наявну літературу, присвячену досліджуваній проблематиці, ми розглянемо застосування великих мовних моделей та графових баз даних. Перша робота [2] досліджує інтеграцію ВММ з графовими базами даних, які виступають як джерело знань, акцентуючи увагу на проблемах коректності ВММ та пропонуючи потенційні шляхи їх вирішення. Друга стаття [3] представляє архітектурні модифікації, спрямовані на забезпечення масштабованості графових баз даних для обробки великих обсягів даних у розподілених обчислювальних середовищах. Це критично важливо для проєктування баз даних, здатних відповідати вимогам сучасних ВММ. Третя стаття [4] зосереджена на оптимізації швидкості виконання запитів до графових баз даних, особливо в контексті зв'язків "багато до багатьох", що є ключовим фактором для забезпечення ефективної роботи систем. Четверта стаття [5] пропонує альтернативний підхід до глибокого навчання, використовуючи виключно реляційні бази даних, що надає цінний досвід у представленні та обробці мовних структур. Цей досвід може бути потенційно корисним для структурування та оптимізації графових баз даних.

Використання графових баз даних для представлення великих мовних моделей пропонує суттєві переваги. Застосування концепції, де нейрони моделі візуалізуються як вершини, а ваги – як ребра, забезпечує можливість детального моделювання взаємозв'язків між параметрами.

Насамперед, це сприяє покращенню інтерпретованості моделі. Відстеження шляхів у графічній структурі дозволяє аналізувати процес прийняття рішень, ідентифікуючи ключові фактори, що вплинули на конкретну відповідь.

Також, аналіз розподілу ваг у графовій базі даних дозволяє виявляти систематичні упередження, що виникають внаслідок незбалансованості тренувальних даних. Можливість ручного коригування ваг або навіть видалення цілих підграфів забезпечує можливість мінімізації негативного впливу таких упереджень та підвищення точності результатів.

Крім того, графова структура забезпечує гнучкість архітектури моделі. Модель може динамічно активовувати певні підграфи для виконання специфічних завдань. Це також відкриває перспективи для персоналізації моделі, дозволяючи адаптувати окремі підграфи до потреб конкретного користувача без необхідності повного перенавчання.

Варто зазначити, що графова база даних дозволяє вести детальну історію змін ваг, що значно економить обчислювальні ресурси порівняно з повним копіюванням ваг.

Інтеграція графових баз даних з великими мовними моделями є перспективним, але складним підходом, що дозволяє підвищити інтерпретованість моделі, забезпечити гнучкість та персоналізувати відповіді. Практична реалізація стикається з обмеженнями, зокрема зниженням швидкості обробки та великими обсягами даних. Незважаючи на це, потенціал інтеграції залишається значним, і потребує подальшого детального аналізу та експериментального дослідження для подолання існуючих викликів.

**Література**

1. Bhattacharjee, A., Moraffah, R., Garland, J., & Liu, H. (2024). *Towards LLM-guided Causal Explainability for Black-box Text Classifiers* (arXiv:2309.13340). arXiv.<https://doi.org/10.48550/arXiv.2309.13340>
2. Besta, M., Gerstenberger, R., Fischer, M., Podstawski, M., Blach, N., Egeli, B., Mitenkov, G., Chlapek, W., Michalewicz, M., Niewiadomski, H., Müller, J., & Hoefler, T. (2023). The Graph Database Interface: Scaling Online Transactional and Analytical Graph Workloads to Hundreds of Thousands of Cores. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 1–18.<https://doi.org/10.1145/3581784.3607068>
3. Moar, C. (2024). *Compressing Language Models using Low-Rank Decomposition and Characterizing the Accuracy—Efficiency Trade-offs*.<https://escholarship.org/uc/item/0t6967h4>
4. Du, L. (2020). *In-Machine-Learning Database: Reimagining Deep Learning with Old-School SQL* (arXiv:2004.05366). arXiv.<https://doi.org/10.48550/arXiv.2004.05366>
5. Yang, L., Chen, H., Li, Z., Ding, X., & Wu, X. (2024). Give us the Facts: Enhancing Large Language Models With Knowledge Graphs for Fact-Aware Language Modeling. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3091–3110. IEEE Transactions on Knowledge and Data Engineering.<https://doi.org/10.1109/TKDE.2024.3360454>