**УДК 004.8:004.932**
**DOI https://doi.org/10.32782/tnv-tech.2025.1.4**

# ОЦІНКА ЕФЕКТИВНОСТІ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ У ЗАДАЧІ ЗАПОВНЕННЯ ГРАФІВ ЗНАНЬ: МЕТОДИ ТА ПЕРСПЕКТИВИ

***Кондрат Р. Я.*** *– кандидат фізико-математичних наук, старший викладач кафедри інформаційних технологій та вищої математики Національного університету біоресурсів і природокористування України «Бережанський агротехнічний інститут»*
*ORCID ID: 0009-0005-1993-4254*

***Білоус Н. М.*** *– старший викладач кафедри інформаційних технологій вищої математики Національного університету біоресурсів і природокористування України «Бережанський агротехнічний інститут»*
*ORCID ID: 0009-0007-9364-6554*

*Автоматизація заповнення графів знань (KG) є ключовим завданням у сфері штучного інтелекту, що знаходить застосування в діалогових системах, пошукових механізмах і аналітичних платформах. У цій роботі досліджується ефективність великих мовних моделей (LLM) у завданні завершення графу знань (KGC) за участю трьох моделей: GPT-4o, GPT-3.5-Turbo-0125 та Mixtral-8x7b-Instruct-v0.1. Дослідження охоплює оцінку продуктивності моделей у Zero-Shot та One-Shot сценаріях, а також аналіз впливу різних підходів до формулювання підказок, включаючи навчання в контексті (ICL) та ланцюжок думок (COT). Використано два спеціалізовані набори даних, що містять як явні, так і неявні зв'язки між сутностями, що дозволяє оцінити здатність моделей до логічних висновків. Результати аналізу отримано за допомогою суворої парадигми, яка вимагає точного збігу передбачених трійок з еталонними, та гнучкої, що допускає часткову відповідність із подальшою постобробкою.*

*Результати демонструють, що LLM можуть бути ефективними у завданнях завершення KG, проте їхня продуктивність значною мірою залежить від якості підказок, наявності прикладів та чіткості формату виводу. Деталізовані підказки без прикладів не завжди сприяють покращенню результатів, а Zero-Shot підхід виявляється менш ефективним порівняно з One-Shot методами. Моделі GPT показують вищу узгодженість із заданими інструкціями, тоді як Mixtral-8x7b іноді додає зайвий пояснювальний текст, що ускладнює його інтеграцію у KG. Незважаючи на досягнуті успіхи, LLM стикаються з обмеженнями у дотриманні формату виводу, розпізнаванні неявних зв'язків та залежності від формулювання підказок. Подальші дослідження мають бути спрямовані на оптимізацію підказок, вдосконалення методів навчання та інтеграцію LLM у більш складні системи KG, що дозволить підвищити точність і ефективність автоматизованого поповнення знань.*

***Ключові слова:*** *великі мовні моделі (LLM), графи знань (KG), завершення графу знань (KGC), навчання в контексті (ICL), ланцюжок думок (COT), Zero-Shot та One-Shot навчання, автоматизована обробка знань.*

***Kondrat R. Ya., Bilous N. M. Evaluation of the effectiveness of large language models in the task of knowledge graph completion: methods and perspectives***

*The automation of knowledge graph (KG) completion is a key task in artificial intelligence, with applications in dialogue systems, search engines, and analytical platforms. This study evaluates the effectiveness of large language models (LLMs) in knowledge graph completion (KGC) by analyzing three models: GPT-4o, GPT-3.5-Turbo-0125, and Mixtral-8x7b-Instruct-v0.1. The research assesses model performance in Zero-Shot and One-Shot scenarios and examines the impact of different prompting strategies, including in-context learning (ICL) and chain-of-thought (COT) reasoning. Two specialized datasets are utilized, containing both explicit and implicit entity relationships, enabling the assessment of models' reasoning capabilities. The analysis employs a strict evaluation paradigm, requiring an exact match between predicted and reference triples, as well as a flexible paradigm, allowing for partial matches and post-processing adjustments.*

*The findings demonstrate that LLMs can be effective in KGC tasks, yet their performance heavily depends on prompt quality, the presence of examples, and output formatting precision. Detailed prompts without examples do not consistently improve results, while Zero-Shot prompting proves less effective than One-Shot approaches. GPT models exhibit greater alignment with given instructions, whereas Mixtral-8x7b tends to include additional explanatory text, making its integration into structured KG systems more challenging. Despite advancements, LLMs still face limitations in output formatting, recognition of implicit relationships, and dependency on prompt formulation. Future research should focus on prompt optimization, refining learning methodologies, and integrating LLMs into more complex KG systems, enhancing the accuracy and efficiency of automated knowledge completion.*

***Key words:*** *large language models (LLM), knowledge graphs (KG), knowledge graph completion (KGC), in-context learning (ICL), chain-of-thought (COT), Zero-Shot and One-Shot learning, automated knowledge processing.*

**Постановка проблеми.** Графи знань (*knowledge graphs*, надалі KG) визначаються як графи даних, призначені для накопичення та передачі знань про реальний світ [1, 2]. Вузли в таких графах відповідають сутностям, що становлять інтерес, а ребра відображають можливі взаємозв'язки між цими сутностями. KG інтегруються у різні системи для покращення можливостей зберігання та обробки інформації. Системи орієнтованого на завдання діалогу (*task-oriented dialogue systems*, надалі TOD) разом із чат-ботами є розмовними агентами, здатними підтримувати діалоги природною мовою. Відмінністю систем TOD є їхня спрямованість на вирішення конкретних завдань користувача у визначених доменах [3, 4].

Попередні дослідження зосереджувалися на розробці вдосконаленої онтології для таких систем, яка включає статичний KG, що дозволяє відображати контекст обговорення та зберігати відповідну інформацію. Використання цього підходу забезпечує низку переваг, серед яких є можливість одночасного ведення кількох розмовних потоків у межах одного діалогу, а також застосування KG для перевірки даних як проксі-механізму. Додатково така система підтримує виконання операцій Створити-Отримати-Оновити-Видалити (*Create-Retrieve-Update-Delete*, надалі CRUD) на KG, що є важливою функціональною складовою для роботи з базами знань. CRUD охоплює чотири основні операції над постійними сховищами, зокрема реляційними або об'єктними базами даних, а також іншими видами баз знань, такими як KG [5, 6].

Оскільки системи TOD виконують широкий спектр завдань, їхнє доповнення базовими, але важливими функціями є доцільним. У цьому контексті завдання «Завершення графу знань» (*knowledge graph completion*, надалі KGC) використовується для побудови KG, а завдання «Міркування в графі знань» (*knowledge graph reasoning*, надалі KGR) – для обробки операцій CRUD. Метою KGC є заповнення відсутньої інформації у KG на основі вхідного тексту або попередніх знань [6].

Сучасні підходи, що базуються на правилах зіставлення шаблонів вхідного тексту, накладають обмеження на автентичність діалогів і ускладнюють адаптацію до нових концепцій за межами визначеної онтології. У зв'язку з цим у літературі досліджувалися методи інтеграції нейронних мереж, зокрема налаштування попередньо навченої моделі BERT для ідентифікації намірів користувача та відповідних об'єктів з вхідного тексту [5]. Впровадження глибокого навчання у систему TOD продемонструвало позитивні результати, проте існуючі обмеження повністю усунути не вдалося.

**Мета дослідження.** У даній роботі вивчається потенціал використання великих мовних моделей (*large language* models, надалі LLM) для розв'язання задач KGC у контексті систем TOD. Літературні джерела [6, 7, 8, 9] відзначають

синергію між KG та LLM, оскільки KG можуть збагачувати LLM, надаючи зовнішні знання для висновків та пояснень, тоді як LLM здатні вирішувати завдання, пов'язані з KG, за допомогою підказок природною мовою. Досліджується ефективність LLM у статичних контекстах KG, використовуючи три відомі моделі: Mixtral-8x7b-Instruct-v0.1 [10], GPT-3.5-Turbo-0125 і GPT-4o. Взаємодія з такими моделями передбачає використання підказок, які формуються як інструкції природною мовою для коректного розпізнавання намірів користувача. Оцінюється здатність цих моделей виконувати поставлене завдання за допомогою різних стилів підказок, включаючи створені людиною та адаптовані до конкретної моделі. Кожна підказка належить до певного рівня, визначеного відповідно до таксономії TELeR [9], що включає методи прямого підказування (*direct prompt*, надалі DP), навчання в контексті (*in-context learning*, надалі ICL) і ланцюга думок (*chain-of-thought*, надалі COT) у парадигмах Zero-Shot та One-Shot.

З метою демонстрації застосування LLM у задачах KG використовуються тестові фрази, отримані під час навчання системи TOD. Створено два набори даних, один із яких містить підвищену складність завдяки тестовим випадкам, що вимагають виконання логічних висновків, неявно закладених у підказках. Такий підхід дозволяє не лише оцінити здатність LLM виконувати завдання, пов'язані з KG, але й дослідити їхню інтеграцію в системи TOD. Крім того, аналізуються показники точності та F1-міри [9], що використовуються для пошуку балансу між точністю і повнотою відповідей, для кожної моделі LLM у різних наборах даних відповідно до строгих і гнучких методів оцінювання:

- Оцінюється продуктивність двох різних LLM: відкритої та комерційної – у KGC, застосовуючи різні стилі підказок та рівні складності. Також використовуються три методи підказування (DP, ICL, COT) у двох контекстах (Zero-Shot і One-Shot), що дозволяє отримати цінну інформацію про надійність моделей.
- Представляються два спеціалізовані набори даних, призначені для оцінки продуктивності LLM у задачі KGC за різного рівня складності.
- Досліджується можливість інтеграції LLM у онтологічно розширену систему TOD, аналізуючи використання тестових фраз, характерних для цього середовища.

Завдання завершення графу знань (KGC) передбачає отримання нової інформації на основі вже наявних знань у KG або введених текстових даних. У літературі пропонуються різні підходи до його розв'язання. Зокрема, Джі С., Пан С. та ін. [11] розглядають методи, що базуються на підсистемах (такі як TransE), а також підходи, що включають обґрунтування шляхів відношення, алгоритми ранжування, навчання з підкріпленням, методи на основі правил (наприклад, KALE) і метареляційне навчання із використанням R-GCN або LSTM. Подібну класифікацію підходів пропонують Жанг Ж. та ін. [7], поділяючи їх на нейронні, символічні та нейронно-символічні методи.

Наведені дослідження зосереджуються на використанні нейронних мереж, логічних правил і математичних операцій для вирішення задачі KGC, однак у більшості випадків глибокий аналіз застосування LLM для цих завдань залишається поза увагою. У цьому контексті Пан С. та ін. [6] розглядають взаємодію між LLM і KG, пропонуючи уніфіковану структуру, що також охоплює KGC. Чжу Й. та ін. [8] експериментують із ChatGPT та GPT-4 у цьому завданні та показують, що хоча вони поступаються спеціалізованим попередньо навченим мовним моделям у парадигмах Zero-Shot та One-Shot, їхні можливості міркування часто є конкурентними або навіть перевершують сучасні моделі. Вей Х. та ін. [12] розглядають

багатоступеневу взаємодію з ChatGPT для вилучення релевантної інформації відповідно до заданої схеми. Додатково Хорашадізаде Х. та ін. [13] вивчають здатність базових моделей, таких як ChatGPT, генерувати KG на основі знань, отриманих під час попереднього навчання, а також текстових даних, наданих у підказках. Отримані результати свідчать про перспективність таких підходів у різних сценаріях використання.

На відміну від вищезазначених робіт, сучасні підходи передбачають розширення кількості текстових вводів, що сприяє підвищенню узагальненості отриманих висновків. Зокрема, досліджується ефективність моделей GPT-3.5-Turbo-0125 та GPT-4o у завданні KGC, подібно до підходу, запропонованого Хорашадізаде Х. та ін. [13]. Додатково розглядається LLM із відкритим кодом – Mixtral-8x7b-Instruct-v0.1, що дозволяє дослідити переваги відкритих моделей, зокрема їхню адаптивність та економічну ефективність у порівнянні з комерційними рішеннями, тому що дослідження можливостей Mixtral для завдань, пов'язаних із KG, є одними з перших у цьому напрямі. Ще однією особливістю є розширення варіантів підказок, що не лише збільшує їхню варіативність, а й підвищує відстежуваність завдяки порівнюванню з таксономією TELeR [9]. Крім цього, розглядається можливість інтеграції LLM із онтологічно розширеною системою TOD, що сприяє покращенню її обробки природної мови та функцій, пов'язаних із KG. Для цієї мети використовуються тестові фрази, отримані з програми навчання TOD, що дозволяє створити два набори даних із різним рівнем складності.

Для глибшого аналізу ефективності моделей оцінюються показники точності та оцінки якості передбачень для обох наборів даних. Використовуються дві парадигми вимірювання: суворе та гнучке оцінювання. Відповідно до суворого підходу, показники обчислюються традиційним способом вилучення тексту, підраховуючи кількість передбачених трійок, що збігаються з еталонними значеннями, з урахуванням їхнього точного форматування. Такий метод дозволяє оцінити здатність моделі точно слідувати підказці та обробляти вхідний текст, що важливо для можливості подальшого використання отриманих результатів у автоматизованих конвеєрах обробки даних. Натомість гнучкий підхід допускає незначні помилки форматування, які можуть бути виправлені на етапах постобробки, а також частково коректні трійки, якщо їх зміст залишається змістовно точним. Це дозволяє отримати позитивну оцінку для моделей, які можуть демонструвати нижчу точність у строгому сенсі, проте потребують менше обчислювальних ресурсів порівняно з більш складними альтернативами.

Моделі GPT потребують додаткового етапу постобробки після генерації виводу. Оскільки ці моделі можуть формувати вихідні дані у форматі JSON, іноді вони автоматично додають до відповіді спеціальні теги. Щоб зберегти ідентичність підказок для всіх моделей і виключити зайві елементи з виводу, застосовується процедура постобробки, що дозволяє коректно оцінити лише змістовну частину відповіді.

Додавання детальних інструкцій до підказок без включення прикладів не завжди покращує результати. Аналіз різних рівнів підказок показує, що збільшення кількості інструкцій без додавання прикладів не приводить до стабільного підвищення продуктивності. Зокрема, підказки третього рівня у суворій оцінці демонструють зниження точності. При застосуванні гнучких метрик ця розбіжність майже зникає. GPT-4o демонструє тенденцію до підвищення продуктивності із кожним рівнем у разі використання рукописних підказок. Водночас перефразовані моделлю підказки можуть спричиняти різке зниження точності. Інші моделі

стабільно погіршують результати на третьому рівні: особливо Mixtral-8x7b, що пояснюється його схильністю відтворювати вхідний текст разом із пояснювальною інформацією.

Як і очікувалося, моделі демонструють найкращу ефективність за умови надання якісних прикладів для орієнтації, що підтверджується літературними джерелами [9, 14]. Проте Mixtral-8x7b інколи включає пояснення у вивід, що ускладнює аналіз і призводить до помилкових висновків. Особливо це проявляється у випадках, коли вхідний текст містить тип класу, відсутній в онтології [10]. Незважаючи на те, що моделі GPT демонструють подібну поведінку значно рідше, LLM усе ще мають значний потенціал для покращення своїх навичок міркування.

Здатність моделей слідувати заданому формату виводу суттєво відрізняється залежно від моделі. Дві метрики оцінювання дозволяють отримати додаткову інформацію про цю здатність:

- Mixtral-8x7b рідко дотримується заданого формату, часто додаючи пояснювальні елементи.
- GPT-3.5-Turbo демонструє мінімальну різницю між суворими та гнучкими показниками.
- GPT-4o зазнає значного зниження продуктивності у разі використання перефразованих підказок, особливо на третьому рівні.

Перспективним є факт, що модель з відкритим вихідним кодом потенційно може покращити свої результати за умови чіткого дотримання системних підказок. GPT-4o демонструє стабільні результати, тоді як GPT-3.5-Turbo іноді показує кращі показники на складніших наборах даних. Водночас попри коливання ефективності на різних рівнях підказок, GPT-4o залишається найсильнішою моделлю в цілому.

**Висновки.** Великі мовні моделі (LLM) демонструють значний потенціал у задачах завершення графу знань (KGC), що відкриває перспективи для їх застосування у сфері штучного інтелекту та автоматизованих систем освіти. Водночас продуктивність цих моделей значною мірою залежить від якості підказок, наявності відповідних прикладів і вибору методу взаємодії з системою. Аналіз показує, що сучасні LLM добре справляються з завданнями, що вимагають явного представлення знань та простих логічних висновків, особливо за умов Zero-Shot та One-Shot навчання.

Попри зростаючі можливості LLM у роботі з KG, моделі все ще мають певні обмеження. Зокрема, вони не завжди точно слідують формату підказки, а також можуть генерувати зайву або неправильно структуровану інформацію, що потребує додаткової постобробки. Певні архітектури більш схильні до генерації пояснювального тексту або додавання зайвих деталей, що ускладнює їхню інтеграцію у формальні системи KG. Іншою важливою проблемою залишається обмежена здатність моделей до неявного міркування, тому коли необхідно враховувати приховані зв'язки між елементами графу знань, результати можуть бути неповними або неточними. Це свідчить про те, що сучасні LLM ще не досягли рівня повноцінного розуміння та обробки складних структурованих знань.

Залежно від специфічних вимог до точності, швидкості та витрат, вибір LLM має бути зваженим. У реальних застосуваннях важливо враховувати і точність моделі, і її можливості в оптимізації обчислювальних витрат. Використання більш легких моделей може бути доцільним у випадках, коли точність не є критичною, тоді як складніші LLM можуть застосовуватися для завдань, що вимагають високої гнучкості та глибокого розуміння контексту.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Hogan A., et al. Knowledge Graphs. Synthesis Lectures on Data. *Semantics, and Knowledge, Morgan & Claypool Publishers.* 2021.

2. Fill H., Fettke P., Keopke J. Conceptual modeling and large language models: Impressions from first experiments with ChatGPT. *Enterpise Modelling & Information Systems Architectures. International Journal of Conceptual Modelling.* 2023. № 18(3). pp. 1-15.

3. Кундос М.Г., Соловей Л.Я. та ін. Ефективність і багатопотоковість паралельних обчислень у системному програмуванні. *Таврійський науковий вісник.* 2024. № 5. С. 60-64.

4. Chen H., Liu X., Yin D., Tang J. A survey on dialogue systems: recent advances and new frontiers. *ACM SIGKDD Explorations News.* 2017. № 19(2). pp. 25-35.

5. Iga V., Silaghi G. Leveraging BERT for natural language understanding of domain-specific knowledge. *25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, 2023, Nancy, France.* pp. 210-215.

6. Pan S., Luo L., Wang Y., Chen C., Wang J., Wu X.: Unifying Large Language Models and knowledge graphs: A roadmap. *CoRR abs/2306.08302*. 2023. https://doi.org/10.48550/ARXIV.2306.08302

7. Zhang J., Chen B., Zhang L., Ke X., Ding H. Neural, symbolic and neuralsymbolic reasoning on knowledge graphs. *AI Open 2*. 2021. pp. 14-35.

8. Zhu Y., et al. LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. *CoRR abs/2305.13168*. 2023. https://doi.org/10.48550/ARXIV.2305.13168

9. Santu S., Feng D. TELLeR: A general taxonomy of LLM prompts for benchmarking complex tasks. *Findings of the ACL: EMNLP 2023, Singapore.* 2023. pp. 14197-14203.

10. Jiang A.Q., et al. Mixtral of Experts. *CoRR abs/2401.04088*. 2024. https://doi.org/10.48550/ARXIV.2401.04088

11. Ji S., Pan S., Cambria E., Marttinen P., Yu P.S. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.* 2022. № 33(2). pp. 494-514.

12. Wei X., et al. ChatIE: zero-shot information extraction via chatting with ChatGPT. *CoRR abs/2302.10205*. 2023. https://doi.org/10.48550/arxiv.2302.10205

13. Khorashadizadeh H., Mihindukulasooriya N., Tiwari S., Groppe J. Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text. *CEURWorkshop Proceedings*. 2023. № 3447, pp. 132-153.

14. Zhao W., et al. A survey of large language models. *CoRR abs/2303.18223*. 2023. https://doi.org/10.48550/ARXIV.2303.18223

## REFERENCES:

1. Hogan A., et al. (2021). Knowledge Graphs. Synthesis Lectures on Data. *Semantics, and Knowledge, Morgan & Claypool Publishers*.

2. Fill H., Fettke P., Keopke J. (2023). Conceptual modeling and large language models: Impressions from first experiments with ChatGPT. *Enterpise Modelling & Information Systems Architectures. International Journal of Conceptual Modelling*, 18 (3), 1-15.

3. Kundos M.H., Solovey L.Ya. (2024). Efektyvnist′ i bahatopotokovist′ paralel′nykh obchyslen′ u systemnomu prohramuvanni [Efficiency and multithreading of parallel computing in system programming]. Tavriys′kyy naukovyy visnyk – Tavria Scientific Bulletin, 5, 60-64 [in Ukrainian].

4. Chen H., Liu X., Yin D., Tang J. (2017). A survey on dialogue systems: recent advances and new frontiers. *ACM SIGKDD Explorations News,* 19 (2), 25-35.

5. Iga V., Silaghi G. (2023). Leveraging BERT for natural language understanding of domain-specific knowledge. *25th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, Nancy, France,* 210-215.

6. Pan S., Luo L., Wang Y., Chen C., Wang J., Wu X. (2023). Unifying Large Language Models and knowledge graphs: A roadmap. *CoRR abs/2306.08302*. https://doi.org/10.48550/ARXIV.2306.08302

7. Zhang J., Chen B., Zhang L., Ke X., Ding H. (2023). Neural, symbolic and neural-symbolic reasoning on knowledge graphs. *AI Open 2*, 14-35.

8. Zhu Y., et al. (2023). LLMs for knowledge graph construction and reasoning: recent capabilities and future opportunities. *CoRR abs/2305.13168*. https://doi.org/10.48550/ARXIV.2305.13168

9. Santu S., Feng D. (2023). TELLeR: A general taxonomy of LLM prompts for benchmarking complex tasks. *Findings of the ACL: EMNLP 2023, Singapore,* 14197-14203.

10. Jiang A.Q., et al. (2024). Mixtral of Experts. *CoRR abs/2401.04088*. https://doi.org/10.48550/ARXIV.2401.04088

11. Ji S., Pan S., Cambria E., Marttinen P., Yu P. (2022). A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans. Neural Networks Learn. Syst.,* 33 (2), 494-514.

12. Wei X., et al. (2023). ChatIE: zero-shot information extraction via chatting with Chat-GPT. *CoRR abs/2302.10205*. https://doi.org/10.48550/arxiv.2302.10205

13. Khorashadizadeh H., Mihindukulasooriya N., Tiwari S., Groppe J. (2023). Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text. *CEURWorkshop Proceedings*, 3447, 132-153.

14. Zhao W., et al. (2023). A survey of large language models. *CoRR abs/2303.18223*. https://doi.org/10.48550/ARXIV.2303.18223