Graph Consistency Rule Mining with LLMs: an Exploratory Study [Extended Abstract]

Hoa Thi Le
Lyon1 University, CNRS Liris
Lyon, France
thi-hoa.le@liris.cnrs.fr

Angela Bonifati
Lyon1 University, CNRS Liris & IUF
Lyon, France
angela.bonifati@univ-lyon1.fr

Andrea Mauri
Lyon1 University, CNRS Liris
Lyon, France
andrea.mauri@univ-lyon1.fr

ABSTRACT

In this extended abstract, we summarize our previously published work (28th EDBT 2025) on leveraging Large Language Models (LLMs) to automatically generate consistency rules for property graphs. Graph data structures are essential for representing complex relationships in various domains, including life sciences, social media, healthcare, finance, security, and planning. With the increasing reliance on graph databases—particularly property graphs—for capturing semantic relationships, ensuring data integrity and quality has become crucial. Traditional methods for maintaining consistency, such as expert-defined rules and data-mined constraints like functional and entity dependencies, face challenges in scalability, adaptability, and comprehensibility. We explore how Large Language Models (LLMs) can be utilized to automatically generate and refine consistency rules for property graphs through guided prompts. Leveraging the reasoning capabilities of LLMs over expressive graph models, we conduct an exploratory empirical study to assess the extent to which LLMs can extract rules that enforce data consistency. Our evaluation spans different real-world datasets and various graph encoding methods. Our results demonstrate that LLMs show promising abilities in extracting consistency rules, primarily identifying schema-based constraints such as primary keys, attribute uniqueness, and label enforcement. Additionally, LLMs occasionally capture more complex patterns, including temporal constraints where certain events cannot occur simultaneously.

VLDB Workshop Reference Format:
Hoa Thi Le, Angela Bonifati, and Andrea Mauri. Graph Consistency Rule Mining with LLMs: an Exploratory Study [Extended Abstract]. VLDB 2025 Workshop: 1st Workshop on New Ideas for Large-Scale Neurosymbolic Learning Systems (LS-NSL).

VLDB Workshop Artifact Availability:
https://github.com/LeHoa98ptit/Rule-mining-with-llms.

1 INTRODUCTION

Graph data has become widespread in various domains [13] such as life sciences, social media, healthcare, finance, security, and planning due to its ability to represent complex relationships between entities. With the growing reliance on graph data, ensuring data integrity and quality has become essential. Graph databases leveraging property graphs have been extensively adopted to capture the semantics of these complex relationships. However, ensuring consistency within large evolving property graphs is challenging. One of the approaches to ensure the quality of the graph is through consistency rules, such as functional dependencies [4] and entity dependencies [3]. These rules help maintain data integrity by enforcing specific constraints and relationships among the entities.

For example, consider a graph representing a social media platform like Twitter, where users, tweets, and hashtags are represented as nodes and edges represent relationships such as mentions, posts, follows, or tags. A consistency rule in this context could enforce temporal constraints—for instance, a retweet can occur only after the original tweet has been posted. Another rule might ensure that users cannot follow themselves or that every tweet must be associated with a valid user who posted it. These rules are crucial for maintaining the logical consistency of the data and preventing anomalies that could affect analytics and user experience.

Traditionally, these rules are either provided by domain experts [11], reflecting business logic related to the data, or mined directly from the data by considering the co-occurrence of elements [8, 14]. However, both approaches have limitations. Expert-defined rules may not cover all edge cases or adapt quickly to new data patterns, while data-mined rules can generate an overwhelming number of constraints, some of which may be redundant, irrelevant, or difficult to understand by the domain expert.

Given the recent advent of Large Language Models (LLMs), many works have started to investigate their capabilities to reason over structured data, both relational and graphs. In the context of graph data, LLMs have shown promising results in basic graph computational tasks on simple labeled graphs [5], graph mining [7], and reasoning [1]. They also enable non-experts to interact with these rich data structures through conversational interfaces [12].

For these reasons, in this paper, we explore how Large Language Models can be used to automatically generate and refine rules for property graphs by guiding them through designed prompts. By leveraging the capabilities of LLMs to understand and reason about graph data structures, we aim to provide an intuitive method to maintain data integrity in graph databases.

In this paper, we contribute with a **exploratory study** to investigate **to what extent** LLMs can **reason** over expressive graph models - such as property graph - to extract rules that can be used to enforce consistency in the data. To this end, we perform an **empirical study** evaluating how different LLMs perform in extracting consistency rules, on various real-world datasets using different graph encoding methods. Our preliminary results show that LLMs have promising capabilities in extracting consistency

Proceedings of the VLDB Endowment. ISSN 2150-8097.

rules, mainly consisting of schema-based constraints (e.g, primary keys, uniqueness of attributes, or forcing specific node or edge labels), but sometimes also extracting constraints related to more complex patterns or considering the temporality of data (e.g., two events cannot happen simultaneously).

2 EXPERIMENTS

Our method encodes graphs into text prompts, guides LLMs to generate interpretable rules, and translates them into executable Cypher queries. We evaluated the ability of LLMs to generate consistency rules using two graph encoding methods (Sliding Attention Window and RAG) and two prompting types (zero-shot and few-shot). We use open-source models `Mixtral` [2] and `LLaMA-3` [10], which can be deployed locally. Full details are provided in our full paper [9]

We conducted our experiments on three real-world property graphs: *WWC20219* (2,468 nodes, 14,799 edges), *Cybersecurity* (953 nodes, 4,838 edges), and *Twitter* (43,325 nodes, 56,493 edges). To evaluate the effectiveness of the consistency rules generated for property graphs we use some ranking measures used in the state-of-the-art of rules mining [6], **support**, **coverage** and **confidence**, and adapted it for property graph.

2.1 Rule Generation

In this section, we present the results in terms of the metrics obtained by the rules generated by the LLMs. Tables 1 report the score after correcting the Cypher queries. The details on how the correction was done are described in Section 2.2

Table 1 shows the results for the `Twitter` dataset. `LLaMA-3` outperforms Zero-shot in terms of support, coverage, and confidence compared to `Mixtral`. Meanwhile, both models demonstrate significant improvement in Few-shot. Specifically, `Mixtral` shows a great improvement with RAG, achieving an average of 100% of coverage and confidence. `LLaMA-3` still dominates the other cases with coverage and confidence values ranging from 70% to 85%.

Regarding the other two datasets, WWC2019 and Cybersecurity, we observe similar trends. The evaluation on the `WWC2019` dataset reveals that `LLaMA-3` generally outperforms `Mixtral` in terms of support, coverage, and confidence, particularly in the Zero-Shot. While `Mixtral` shows lower quantitative performance, it tends to generate more complex and nuanced rules. In Few-Shot, both models improve, but `LLaMA-3` maintains stronger overall scores. For the `Cybersecurity` dataset, `LLaMA-3` again performs better with Sliding Window Attention, whereas `Mixtral` achieves higher support under the RAG method, though with lower coverage and confidence. Few-Shot prompting improves results for both models, with `LLaMA-3` remaining more consistent across encoding methods.

	Sliding Window Attention				RAG			
	#rules	Supp%	Cov%	Conf%	#rules	Supp%	Cov%	Conf%
Zero-shot								
Llama-3	8	12177	72.27	86.14	8	981	70.62	78.75
Mixtral	10	10789	81.20	81.20	7	7698	67.3	76
Few-shot								
Llama-3	7	25201	85.72	85.72	9	8994	71.34	77.78
Mixtral	7	15262	78.79	83.25	8	11593	100	100

Table 1: Support, coverage and confidence score for the `Twitter` dataset with Zero-Shot and Few-Shot Prompts

2.2 Cypher Generation

Model	Sliding Window Attention		RAG	
	Zero-shot	Few-shot	Zero-shot	Few-shot
WWC-2019				
Llama-3	11/12	7/8	7/7	5/6
Mixtral	8/9	7/8	5/6	4/5
Cybersecurity				
Llama-3	8/10	7/9	6/7	7/7
Mixtral	7/10	4/5	4/6	4/5
Twitter				
Llama-3	7/8	5/7	7/8	8/9
Mixtral	9/10	6/7	6/7	8/8

Table 2: Number of correctly generated Cypher queries

In this section, we discuss the performances of the model in generating the Cypher queries related to the rules. We consider a query not correct if it has syntax errors or if its formulation does not match the data model. As shown in Table 2, While both LLMs achieve over 70% query correctness, three main error types were observed: (1) incorrect relationship direction, (2) hallucinated or non-existent properties, and (3) syntax errors. In addition to these error categories, another factor contributing to the decrease in LLM performance is the generation of **inaccurate rules** (i.e., the rule itself is not correct). To ensure a fair evaluation of the LLM's ability to generate consistency rules, we corrected the queries in case of syntax errors or wrong edge directions, but we left them as they were the queries with additional non-existing properties, because those errors corresponded to hallucination at rule generation level, rather than the translation to Cypher.

3 CONCLUSIONS

In our study, we initially aimed to extract specific GFD and GED rules. However, we observed that the LLMs struggled to distinguish between these concepts effectively. In general, for all the datasets, the extracted rules seem to relate to the schema of the graph (e.g., enforcing that nodes are connected with edges having specific labels, or specifying some values for the properties). `LLaMA-3` generates simple rules, while `Mixtral` produces more complex but harder-to-apply rules. The sliding window method is more effective than RAG because RAG has issues with context retrieval. Few-Shot prompting only improves results on the Twitter dataset. Manual correction is still needed, but automation may be possible in the future, and the pipeline could be made interactive to allow domain experts to easily refine the rules.

ACKNOWLEDGMENTS

The work was partially funded by the Data4Health ANR/CPJ Lyon 1 and the ANR-24-CE25-6501 CITADEL grants.

REFERENCES

[1] Francesco Cambria, Francesco Invernici, Anna Bernasconi, and Stefano Ceri. 2024. MINE GRAPH RULE: A New Cypher-like Operator for Mining Association Rules on Property Graphs. *arXiv preprint arXiv:2406.19106* (2024). https://doi.org/10.48550/arXiv.2406.19106 Submitted on 27 Jun 2024.

[2] Albert Q. Jiang et al. 2024. Mixtral of Experts. *arXiv preprint arXiv:2401.04088* (2024). https://doi.org/10.48550/arXiv.2401.04088 Both the base and instruct models are released under the Apache 2.0 license.

[3] Wenfei Fan and Ping Lu. 2017. Dependencies for Graphs. In *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (Chicago, Illinois, USA) *(PODS '17)*. Association for Computing Machinery, New York, NY, USA, 403–416. https://doi.org/10.1145/3034786.3056114

[4] Wenfei Fan, Yinghui Wu, and Jingbo Xu. 2016. Functional Dependencies for Graphs. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 1843–1857. https://doi.org/10.1145/2882903.2915232

[5] Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a Graph: Encoding Graphs for Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR) 2024*. International Conference on Learning Representations, Vienna, Austria. https://iclr.cc/Conferences/2024/AuthorGuide

[6] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *Proceedings of the 22nd international conference on World Wide Web*. 413–422.

[7] Jiayan Guo, Lun Du, Hengyu Liu, Mengyu Zhou, Xinyi He, and Shi Han. 2023. GPT4Graph: Can Large Language Models Understand Graph Structured Data? An Empirical Evaluation and Benchmarking. *arXiv preprint arXiv:2305.15066* (2023). https://doi.org/10.48550/arXiv.2305.15066 Version 2.

[8] Jonathan Lajus, Luis Galárraga, and Fabian Suchanek. 2020. Fast and Exact Rule Mining with AMIE 3. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings* (Published: 31 May 2020). Springer, 36–52. https://doi.org/10.1007/978-3-030-49461-2_3

[9] Hoa Thi Le, Angela Bonifati, and Andrea Mauri. 2025. Graph Consistency Rule Mining with LLMs: an Exploratory Study. In *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025*, Alkis Simitsis, Bettina Kemme, Anna Queralt, Oscar Romero, and Petar Jovanovic (Eds.). OpenProceedings.org, 748–754. https://doi.org/10.48786/EDBT.2025.60

[10] Meta AI. 2024. Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date. https://ai.meta.com/blog/meta-llama-3/

[11] Victoria Nebot and Rafael Berlanga. 2012. Finding association rules in semantic web data. *Knowledge-Based Systems* 25, 1 (2012), 51–62. https://doi.org/10.1016/j.knosys.2011.05.009 Special Issue on New Trends in Data Mining.

[12] Yun Peng, Sen Lin, Qian Chen, Shaowei Wang, Lyu Xu, Xiaojun Ren, Yafei Li, and Jianliang Xu. 2024. ChatGraph: Chat with Your Graphs . In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 5445–5448. https://doi.org/10.1109/ICDE60146.2024.00424

[13] Yuanyuan Tian. 2022. The World of Graph Databases from An Industry Perspective. arXiv:2211.13170 [cs.DB] https://arxiv.org/abs/2211.13170

[14] Zezhong Xu, Peng Ye, Hui Chen, Meng Zhao, Huajun Chen, and Wen" Zhang. 2022. Ruleformer: Context-aware Rule Mining over Knowledge Graph. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2551–2560. https://aclanthology.org/2022.coling-1.225