Effective Generation of Structured Data using LLMs

1 Context and motivation

Rather than generating unstructured text, LLMs are being extensively used for producing structured output that adheres to a predefined schema [8]. This schema can be defined in various formalisms ranging from regular expressions to more expressive languages like JSON Schema. Such generation is crucial in different contexts. For example, LLM agents interact with other agents and with external tools through the A2A [1] and MCP [2] protocols using messages specified by JSON Schema /.

Several techniques have been proposed for using LLMs to produce JSON data from JSON Schema [9, 10]. Most of them rely on constrained decoding, a popular approach that operates directly during the LLM decoding process by restricting the set of valid tokens so that the generated output conforms to a given grammar or structural specification. To support JSON Schema, existing techniques like XGrammar[7] and Outlines [11] rely on a partial translation of the input schema S into a more or less equivalent grammar G that will be used for the constrained decoding. Since JSON Schema is expressive and S is expected to be large and complex, such techniques relying usually fail to guarantee correctness, and suffer strong limitations.

While JSON Schema has been popularized for describing data exchanged through APIs and agentic workflows, recently graph schema languages like PG-Schema [3] are emerging and being adopted in major database systems and for describing graphs used by agents to augment LLMs with retrieval features [6]. Leveraging LLMs to generating graph data starting from graph schemas is an appealing direction that has not yet been investigated and which could create momentum.

2 Objectives

The aim of this doctoral project is to investigate the generation of structured data using LLMs by considering two important data models: JSON and property graphs. While LLM-based generation of JSON data starting from JSON Schema has been extensively studied, existing solutions suffer severe limitations both in terms of robustness and effectiveness. The first objective of the thesis is to address these limitations by leveraging constrained decoding techniques in order to fully support JSON Schema while ensuring efficiency which is paramount when using LLMs. One possible direction is to extend constrained decoding techniques by allowing them to consider larger fragments of JSON Schema. Another direction would be to investigate hybrid approaches combining symbolic solutions [4] with LLMs [9].

The second objective is to study LLM-based generation for property graphs starting from PG-Schema [3] specifications. Despite being dedicated to graph data, this schema language shares some common ground with JSON Schema when it comes to specify structured information represented on the nodes and the edges of the graphs being described. Still, the complexity of the problem is expected to be high due to the nature of the data model which connects nodes following some patterns and probability distributions that need to be captured and reflected during generation. An interesting direction to investigate is to leverage the use of probabilistic generations techniques like GMark [5] with LLMs.

3 Prerequisite

The applicants should have a background in computer science with proficiency in databases, machine learning and artificial intelligence. The candidate will be involved in a research project about effective LLM-guided generation involving researchers from Italy and from Germany following an intensive interaction. Fluency in English is, thus, required.

4 Contact information

The PhD project will be held under the joint supervision of Dr. Mohamed-Amine Baazizi (`mohamed-amine.baazizi@lip6.fr`) and Pr. Dario Colazzo (`dario.colazzo@lamsade.dauphine.fr`) following an extensive interaction between two highly ranked research institutions, namely, the LIP6 lab at Sorbonne University and the LAMSADE lab at PSL University.

References

[1] Agent2agent (a2a) protocol, 2026.

[2] Model context protocol (mcp), 2026.

[3] Renzo Angles, Angela Bonifati, Stefania Dumbrava, George Fletcher, Alastair Green, Jan Hidders, Bei Li, Leonid Libkin, Victor Marsault, Wim Martens, Filip Murlak, Stefan Plantikow, Ognjen Savkovic, Michael Schmidt, Juan Sequeda, Slawek Staworko, Dominik Tomaszuk, Hannes Voigt, Domagoj Vrgoc, Mingxi Wu, and Dusan Zivkovic. Pg-schema: Schemas for property graphs. *Proc. ACM Manag. Data*, 1(2), June 2023.

[4] Lyes Attouche, Mohamed Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. Witness Generation for JSON Schema. *Proc. VLDB Endow.*, 15(13):4002–4014, 2022.

[5] Guillaume Bagan, Angela Bonifati, Radu Ciucanu, George H. L. Fletcher, Aurélien Lemay, and Nicky Advokaat. gmark: Schema-driven generation of graphs and queries. *IEEE Trans. Knowl. Data Eng.*, 29(4):856–869, 2017.

[6] Mariam Barry, Gaëtan Caillaut, Pierre Halftermeyer, Raheel Qader, Mehdi Mouayad, Fabrice Le Deit, Dimitri Cariolaro, and Joseph Gesnouin. Graphrag: leveraging graph-based efficiency to minimize hallucinations in llm-driven rag for finance data. In *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pages 54–65, 2025.

[7] Yixin Dong, Charlie F Ruan, Yaxing Cai, Ruihang Lai, Ziyi Xu, Yilong Zhao, and Tianqi Chen. Xgrammar: Flexible and efficient structured generation engine for large language models. *Proceedings of Machine Learning and Systems 7*, 2024.

[8] Michael Xieyang Liu, Frederick Liu, Alexander J. Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J. Cai. "we need structured output": Towards user-centered constraints on large language model output. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery.

[9] Yu Liu, Duantengchuan Li, Kaili Wang, Zhuoran Xiong, Fobo Shi, Jian Wang, Bing Li, and Bo Hang. Are llms good at structured outputs? a benchmark for evaluating structured output capabilities in llms. *Information Processing & Management*, 61(5):103809, 2024.

[10] Darren Yow-Bang Wang, Zhengyuan Shen, Soumya Smruti Mishra, Zhichao Xu, Yifei Teng, and Haibo Ding. Slot: Structuring the output of large language models, 2025.

[11] Brandon T. Willard and RÃĺmi Louf. Efficient guided generation for large language models. *https://arxiv.org/abs/2307.09702*, 2023.