What can LLMs do to generate synthetic semistructured data?

Automatic generation of synthetic data is essential for testing purposes but also for training ML models and benchmarking systems [9, 6]. In many situations, data is represented in JSON and its format specified in JSON Schema, a logical language for describing complex JSON documents, and many solutions for generating data from JSON Schema exist [4, 5, 7]. However, these solutions suffer from serious problems which hinder their adoption: the generated data is random and not always valid w.r.t. the input schema and it is not realistic enough to reflect the domain intended by the user specifying the schema.

Recently, LLMs have been employed for solving several data management problems like generating SQL queries from user input or performing data wrangling based on user needs but only one attempt for generating JSON data from JSON Schema specifications exists [2] without being satisfactory because it fails at dealing with large schema specifications and does not understand logical operators. Other models like [1] dedicated to code completion are too specialized to deal with JSON Schema.

The workplan of this project as follows:

1. study the limitation of [2] and survey solutions for synthetic data generation using LLMs
2. build a post-processing layer for validating the output of [2] using an external validator
3. suggest a solution based on combining [2] with a fixing strategy when the generated data is not valid
4. study the possibility of fine tuning [2] or [1] by using inputs obtained from an ongoing effort for building a deterministic generator [3].

Contact information: `mohamed-amine.baazizi@lip6.fr` and `lyes.attouche@dauphine.psl.eu`

References

[1] Hugging Face/replit, 2024. https://huggingface.co/replit/replit-code-v1-3b.

[2] Jsonformer: A Bulletproof Way to Generate Structured JSON from Language Models., 2024. https://github.com/1rgs/jsonformer.

[3] Lyes Attouche, Mohamed-Amine Baazizi, and Dario Colazzo. Overview and perspectives for optimistic json schema witness generation. In *BDA*, 2023.

[4] Lyes Attouche, Mohamed-Amine Baazizi, Dario Colazzo, Giorgio Ghelli, Carlo Sartiani, and Stefanie Scherzinger. Witness generation for JSON schema. *Proc. VLDB Endow.*, 2023.

[5] Clara Benac Earle, Lars-Åke Fredlund, Ángel Herranz, and Julio Mariño. Jsongen: a quickcheck based library for testing json web services. In *Proceedings of the Thirteenth ACM SIGPLAN workshop on Erlang*, pages 33–41, 2014.

[6] Angela Bonifati, Irena Holubová, Arnau Prat-Pérez, and Sherif Sakr. Graph generators: State of the art and open challenges. *ACM Comput. Surv.*, 53(2):36:1–36:30, 2021.

[7] Hugo André Coelho Cardoso and José Carlos Ramalho. Synthetic data generation from json schemas. In *11th Symposium on Languages, Applications and Technologies (SLATE 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[8] Kim et al. M2bench: A database benchmark for multi-model analytic workloads. In *VLDB, to appear*, 2023.

[9] Ciprian Paduraru and Marius-Constantin Melemciuc. An automatic test data generation tool using machine learning. In *ICSOFT*, pages 506–515, 2018.