Querying Databases with Natural Language: The use of Large Language Models for Text-to-SQL tasks

Eduardo Roger S. Nascimento[1], **Marco Antonio Casanova**[1]

[1] Department of Informatics
Pontificia Universidade Catolica do Rio de Janeiro (PUC-Rio)
Rio de Janeiro – RJ – Brazil

{enascimento,casanova}@inf.puc-rio.br

Abstract. *The Text-to-SQL task involves generating SQL queries based on a given relational database and a Natural Language (NL) question. Although Large Language Models (LLMs) show good performance on well-known benchmarks, they are evaluated on databases with simpler schemas. This dissertation first evaluates their effectiveness on a complex and openly available database (Mondial) using GPT-3.5 and GPT-4. The results indicate that LLM-based models perform poorly and struggle with schema linking and joins. To improve accuracy, this work proposes the use of LLM-friendly views and data descriptions. A second experiment on a real-world database confirms that this approach enhances the accuracy of the Text-to-SQL task.*

Resumo. *A tarefa de Texto-para-SQL envolve a geração de consultas SQL com base em um banco de dados relacional e uma pergunta em Linguagem Natural (LN). Embora os Modelos de Linguagem Grandes (LLMs) apresentem bom desempenho em benchmarks conhecidos, eles são avaliados em bancos de dados com esquemas mais simples. Esta dissertação avalia inicialmente sua eficácia em um banco de dados complexo e disponível publicamente (Mondial) utilizando GPT-3.5 e GPT-4. Os resultados indicam que os modelos baseados em LLM têm desempenho inferior e dificuldades com a vinculação de esquemas e joins. Para melhorar a precisão, este trabalho propõe o uso de views e descrições de dados amigáveis para LLMs. Um segundo experimento, em um banco de dados do mundo real, confirma que essa abordagem aumenta a precisão na tarefa de Texto-para-SQL.*

1. Data from the Defended Dissertation

- Date of Defense: 04/04/2024
- Graduate Program: Programa de Pós–Graduação em Informática da Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
- Category: Master's degree
- Committee Members: Melissa Lemos Cavaliére (Pontifícia Universidade Católica do Rio de Janeiro), Luiz André Portes Paes Leme (Universidade Federal Fluminense)

2. Context e problem

Text-to-SQL tasks involve generating SQL queries from natural language (NL) sentences [Quamar et al. 2022]. Due to the ability of Large Language Models (LLMs) to understand

natural language and generate high-quality text [Singh 2023], the use of these models in text-to-SQL tasks has gained popularity, achieving considerable success over well-known benchmarks [Yu et al. 2018, Li et al. 2023].

However, these benchmarks are biased towards databases with simple schemas and NL questions aligned with the schema vocabulary. In real-world databases, the schema is often large and complex, and uses abbreviated and ambiguous table and column names, which poses challenges to the text-to-SQL task.

Motivated by the discrepancy between the databases used in benchmarks and those observed in real-world scenarios, this dissertation investigates the performance of *LLM-based text-to-SQL* strategies when applied to complex, real-world databases.

3. Goals

The goal of this dissertation is to assess the performance of LLM-based text-to-SQL strategies on complex databases in two scenarios: (1) the *open scenario* is based on an openly available database [May 1999], which has a large and complex schema and a schema vocabulary close to the terms of the user's questions; (2) the *industrial scenario* is based on a proprietary database that stores data about the integrity management of an oil company industrial assets, whose schema vocabulary is different from that of the user's NL questions.

4. Contribution

The first contribution of this dissertation is to show that, while some LLM-based text-to-SQL strategies showed promising results in benchmarks such as Spider, their performance decreases considerably when applied to complex databases. The second contribution of this dissertation is to demonstrate that LLM-based text-to-SQL strategies can be improved with *LLM-friendly views* and data sample descriptions. The experiments with the Mondial database are available at a Github repository[1].

5. Advancement in the state of the art

Despite the success of LLM-based strategies in the text-to-SQL task, such as [Pourreza and Rafiei 2023], [Dong et al. 2023], and [Gao et al. 2023], on benchmarks like Spider [Yu et al. 2018], these benchmarks typically feature simple and small databases The BIRD dataset aims at bridging the gap between text-to-SQL research and real-world applications, but still lacks databases with large schemas [Li et al. 2023]. This dissertation highlights that LLM performance tends to decline in scenarios involving complex, real-world databases and benchmarks with more challenging queries. It also demonstrates that accuracy can be enhanced by leveraging views [Groff and Weinberg 1999] that adopt LLM-friendly names and that decrease the number of joins during SQL generation. These views help reduce SQL query complexity and facilitate the identification of columns and tables by the LLM during SQL generation. Furthermore, there is potential for further improvements using techniques such as Retrieval Augmented Generation (RAG) [Guo et al. 2023] and fine-tuning an LLM locally [Dettmers et al. 2023] by training the model with data from these more complex schemas.

[1] `https://github.com/dudursn/text\_to\_sql\_chatgpt\_real\_world/`

6. Summary of the Solution

Using both GPT-3.5 and GPT-4 models from OpenAI [OpenAI 2024], this dissertation initially tested selected strategies based on prompt engineering [Saravia 2022] from the Spider benchmark, including DIN-SQL [Pourreza and Rafiei 2023] and C3 [Dong et al. 2023], and a combination of both, called *C3-DIN*, introduced in the dissertation. It also tested SQLQueryChain, SQLDatabaseSequentialChain, and SQLAgent, offered by the LangChain framework [Langchain 2024]. This framework enables including samples of database instances, a feature also tested in this dissertation.

In the industrial scenario, to improve the performance of LLM-based text-to-SQL strategies, this dissertation proposed using views, that map fragments of the database schema to terms closely aligned with those users frequently adopt and pre-define commonly used joins (*LLM-friendly views*) and providing descriptions of the database values to capture the data semantics. The dissertation tested a text-to-SQL implementation based on LangChain's SQLQueryChain that performed well and had a much lower cost than the other strategies tested. Figure 1 illustrates the prompt implemented, where: (A) contains instructions for the LLM; (B) defines the output format; (C) partly illustrates how the `maintenance_order` table is passed to the LLM as a `CREATE TABLE` statement; (D) shows 3 data sample descriptions from the `maintenance_order`; and (E) passes the NL question.

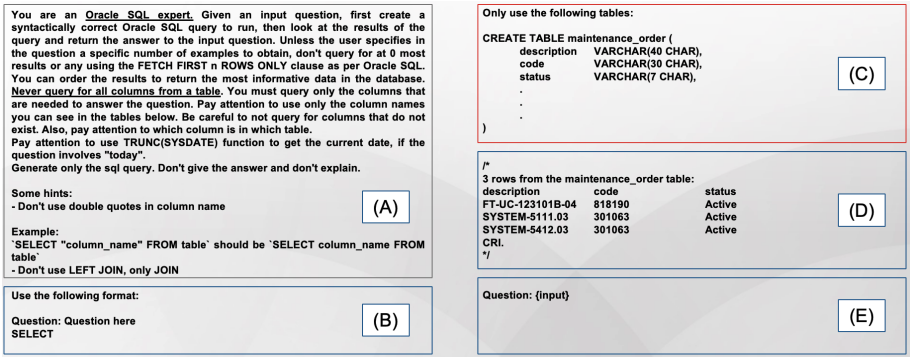

Figura 1. SQLQueryChain's prompt with some tips used in the experiments.

7. Evaluation

Despite the availability of the benchmark datasets for the text-to-SQL task [7, 14, 15], two benchmarks were created, one using the Mondial database and another based on the real-world industrial database mentioned earlier. Each benchmark consists of 100 natural language sentences and their corresponding ground truth SQL queries, categorized as simple, medium and complex. Accuracy, token usage, and cost were the evaluated metrics. Among the strategies tested, SQLQueryChain with samples using GPT-4 showed good performance at a low cost and runtime, while C3 with GPT-4 achieved the best overall accuracy, but incurred higher costs, higher number of tokens, and longer runtime. Results from the Mondial benchmark indicated lower performance due to challenges such as schema linking and joins, complex data semantics, and discrepancies between terms used by the user and the schema vocabulary. In the industrial scenario, renaming tables and columns and introducing new columns to reduce joins improved accuracy in LLM-based text-to-SQL strategies.

8. Scientific and technical production and awards

8.1. Scientific Production

- Pinheiro, J.; Victorio, W.; Nascimento, E. R.; Seabra, A.; Izquierdo, Y.; García, G.; Coelho, G.; Lemos, M.; Leme, L.; Furtado, A. and Casanova, M. (2023). **On the Construction of Database Interfaces Based on Large Language Models**. In *Proceedings of the 19th International Conference on Web Information Systems and Technologies* - WEBIST; ISBN 978-989-758-672-9; ISSN 2184-3252, SciTePress, pages 373-380. DOI: 10.5220/0012204000003584.
- Nascimento, E. R., Garcia, G. M., Victorio, W. Z., Lemos, M., Izquierdo, Y. T., Garcia, R. L., Leme, L. A. P., Casanova, M. A. **A family of natural language interfaces for databases based on chatgpt and langchain**. In: *Proceedings of the 42nd International Conference on Conceptual Modeling* – Posters&Demos. Lisbon, Portugal
- Nascimento, E.; García, G.; Feijó, L.; Victorio, W.; Izquierdo, Y.; R. de Oliveira, A.; Coelho, G.; Lemos, M.; Garcia, R.; Leme, L. and Casanova, M. (2024). **Text-to-SQL Meets the Real-World**. In *Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS*; ISBN 978-989-758-692-7; ISSN 2184-4992, SciTePress, pages 61-72. DOI: 10.5220/0012555200003690 – QUALIS A3.
- R. Nascimento, E.; T. Izquierdo, Y.; García, G.; Coelho, G.; Feijó, L.; Lemos, M.; Leme, L. and Casanova, M. (2024). **My Database User Is a Large Language Model**. In *Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS*; ISBN 978-989-758-692-7; ISSN 2184-4992, SciTePress, pages 800-806. DOI: 10.5220/0012697700003690 – QUALIS A3.

8.2. Awards

- Best Student Paper Candidate Certificate for the paper entitled *"Text-to-SQL Meets the Real-World"* presented at the 26th International Conference on Enterprise Information Systems (ICEIS), and selected for submission to a journal.

8.3. Extensions of the Work

- Coelho, G.M.C., Nascimento, E.R.S., Izquierdo, Y.T., García, G.M., Feijó, L., Lemos, M., Garcia, R.L.S., de Oliveira, A.R., Pinheiro, J.P.V., Casanova, M.A. (2024). **Improving the Accuracy of Text-to-SQL Tools Based on Large Language Models for Real-World Relational Databases**. In: *Strauss, C., Amagasa, T., Manco, G., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Database and Expert Systems Applications*. DEXA 2024. Lecture Notes in Computer Science, vol 14910. Springer, Cham.
- Oliveira, A.R., Nascimento, E.R.S., Coelho, G.M.C., Avila, C.V., Feijó, L., Izquierdo, Y.T., García, G.M., Pinheiro, J.P.V., Leme, L.P.P., Lemos, M., Casanova, M.A. (2024). **Small, Medium, and Large Language Models for Text-to-SQL**. (accepted to ER 2024).
- Nascimento, E.R.S., Avila, C.V., Izquierdo, Y.T., García, G.M., Feijó, L., Coelho, G.M.C., Pinheiro, J.P.V., Leme, L.P.P., Lemos, M., Casanova, M.A. (2024). **On the Text-to-SQL Task for Complex Natural Language Questions Supported by Keyword Query Processing**. (in preparation)

ACKNOWLEDGEMENTS

The authors would like to thank Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Tecgraf Institute and Petrobras for their support and resources that were instrumental in carrying out this research. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

Referências

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Available at: https://arxiv.org/abs/2305.14314.

Dong, X., Zhang, C., Ge, Y., Mao, Y., Gao, Y., lu Chen, Lin, J., and Lou, D. (2023). C3 zero-shot text-to-sql with chatgpt. Available at: https://arxiv.org/abs/2307.07306.

Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., and Zhou, J. (2023). Text-to-sql empowered by large language models a benchmark evaluation. Available at: https://arxiv.org/abs/2308.15363.

Groff, J. R. and Weinberg, P. N. (1999). *SQL: The Complete Reference*. Osborne/McGraw-Hill.

Guo, C., Tian, Z., Tang, J., Li, S., Wen, Z., Wang, K., and Wang, T. (2023). Retrieval-augmented gpt-3.5-based text-to-sql framework with sample-aware prompting and dynamic revision chain. Available at: https://arxiv.org/abs/2307.05074.

Langchain (2024). Langchain is a framework for developing applications powered by language models. Available at: https://python.langchain.com/docs/get_started/introduction.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Cao, R., Geng, R., Huo, N., Zhou, X., Ma, C., Li, G., Chang, K. C. C., Huang, F., Cheng, R., and Li, Y. (2023). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. Available at: https://arxiv.org/abs/2305.03111.

May, W. (1999). Information extraction and integration with FLORID: The MONDIAL case study. Technical Report 131, Universität Freiburg, Institut für Informatik. Available at: http://www.dbis.informatik.uni-goettingen.de/Mondial.

OpenAI (2024). Openai blog. Available at: https://openai.com/blog/new-embedding-models-and-api-updates.

Pourreza, M. and Rafiei, D. (2023). Din-sql: Decomposed in-context learning of text-to-sql with self-correction. Available at: https://arxiv.org/abs/2304.11015.

Quamar, A., Efthymiou, V., Lei, C., and Özcan, F. (2022). Natural language interfaces to data. *Foundations and Trends in Databases*, 11(4):319–414.

Saravia, E. (2022). Prompt Engineering Guide. Available at: https://github.com/dair-ai/Prompt-Engineering-Guide.

Singh, A. (2023). Large language models: A guide on its benefits, use cases, and types. Available at: https://yellow.ai/blog/large-language-models.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., and Radev, D. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics. Available at: https://aclanthology.org/D18-1425.