Exploring Temporal Text-to-SQL Challenges in Brazilian Portuguese: Lessons from Educational Data*

Karina de Carvalho Fróes[1], **Kelly Rosa Braghetto**[1]

[1]Institute of Mathematics and Statistics, University of São Paulo

{karina.froes,kellyrb}@ime.usp.br

Abstract. *Recent advances in natural language processing have enabled translating natural language into SQL, but challenges remain in multilingual and temporal contexts. This short paper presents a beginner-level exploratory analysis of a prompt engineering strategy for text-to-SQL generation over Brazilian Portuguese educational data. Through 10 representative examples reflecting how real users might ask questions about open government data, we show how language variability, implicit temporal references, and mismatched expectations affect SQL generation and the reliability of standard evaluation metrics. This work contributes to historical data querying research and highlights persistent challenges for multilingual text-to-SQL systems.*

1. Introduction

Advances in Natural Language Interfaces to Databases (NLIDB), or Text-to-SQL techniques, have enabled users to query structured data using natural language, reducing the need for technical expertise and increasing accessibility. Recent progress—driven especially by large language models (LLMs)—has improved performance in controlled benchmarks. However, challenges remain when these systems are applied to real-world scenarios, particularly in multilingual and temporal contexts.

Temporal analysis is critical in many domains, from business and healthcare to public policy and education. In the educational context, stakeholders often need to compare data across school years, identify trends, or evaluate policy impacts. Yet, widely used benchmarks such as Spider [Yu et al. 2018], WikiSQL [Zhong et al. 2017], and BIRD [Li et al. 2023] rarely cover temporally grounded questions or reflect the linguistic diversity and ambiguity found in real user inputs—especially in non-English languages.

This short paper presents a preliminary analysis focused on the outputs of a prompt engineering strategy based on the DIN-SQL model [Pourreza and Rafiei 2024] with GPT-4, applied to Brazilian Portuguese questions over the *Censo Escolar*—an open, annually updated government dataset on Brazilian education. We examine 10 representative examples that illustrate how users with varying technical backgrounds might naturally ask questions about historical data. By analyzing the resulting SQL queries, we aim to uncover how linguistic ambiguity, temporal references and mismatched expectations can affect SQL quality and distort the reliability of standard evaluation metrics. By addressing these challenges, we seek to contribute to the development of more robust Text-to-SQL systems for multilingual and temporally grounded scenarios.

*Funding: This work was supported by São Paulo Research Foundation (FAPESP) [grants #2023/00779-0 and #2023/18026-8]; and the National Council for Scientific and Technological Development (CNPq) [grant #420623/2023-0].

2. Conceptualization

Text-to-SQL Models The task of Text-to-SQL consists of translating questions written in natural language into structured queries in SQL, using database schema information — such as table and column names — to produce executable statements [Visperas et al. 2023]. Over time, two main approaches have emerged: rule-based and machine learning-based methods. Rule-based models were among the earliest proposed and can be effective in narrow domains [Őzcan et al. 2020]. However, they do not scale well in the face of the ambiguity and variability of natural language. Machine learning-based approaches, particularly those leveraging attention architectures such as Transformers [Vaswani 2017], have become state-of-the-art due to their ability to learn complex semantic mappings between text and SQL from large annotated datasets.

More recently, large language models (LLMs) such as GPT, T5 and PaLM have been explored using prompting strategies, few-shot learning, and modular task decomposition, showing promising results [Sun et al. 2023]— especially on benchmarks like Spider and BIRD. These models demonstrate advances not only in accuracy metrics but also in their generalization capacity to handle complex queries and heterogeneous databases — a context in which this research is situated.

Temporal Data In relational databases, temporal data captures how values change over time. This is often modeled using time-specific columns or dedicated temporal tables, where each row includes start and end timestamps to track data validity [Levene et al. 1999]. In this work, we consider temporal data as any value that conveys time-related information—explicitly (e.g., a `year` column) or implicitly (e.g., "after November" or "first trimester"). This includes structured types like `date` and `datetime`, as well as temporal expressions such as "month," "June," or "last year."

3. Related Work

Recent works on text-to-SQL have moved from rule-based and neural models [Xu et al. 2017, Wang et al. 2019] to large language models (LLMs), which currently define the state of the art. Among these, DIN-SQL [Pourreza and Rafiei 2024] and DAIL-SQL [Gao et al. 2024] use prompting and decoding constraints to better handle complex queries, while SQL-PaLM [Sun et al. 2023] demonstrates that few-shot prompting with code-pretrained LLMs can generalize well to unseen databases. However, most of these approaches are benchmarked on datasets like Spider and BIRD, with limited temporal diversity and mostly English-language content.

In contrast, our work explores LLM-based SQL generation over real-world historical educational data in Portuguese. We make a brief demonstration of the challenges of linguistic variation, schema complexity, and semantic ambiguity that arise in this setting, even when relying on a consistent model such as DIN-SQL, adapted to our setting.

4. Methodology

4.1. Dataset and Experimental Design

The Brazilian Censo Escolar [1] is the most comprehensive national survey on basic education, collecting yearly data from public and private schools across the country. It provides

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar

detailed and structured historical records that have been publicly available from its inception through 2024. The data is organized into four main relational tables:

- **Escola:** information on school location, administrative dependency, infrastructure, pedagogical staff, and general student characteristics.
- **Turma:** class composition, teaching modality, scheduling, and additional attributes.
- **Matricula:** details on enrolled students, including age, educational level, and whether they have special needs.
- **Docente:** teacher-related data, such as educational background and disciplines.

To explore how large language models handle SQL generation over real-world and temporally rich data, we manually created a reference set of natural language questions and corresponding SQL queries. These were built through careful analysis of the schema and the types of information that are relevant for educational policy and research. From this initial dataset, we selected ten representative questions with varying levels of linguistic complexity and temporal structure to serve as test cases. The goal was to assess the model's behavior before applying more advanced prompt engineering techniques.

We selected DIN-SQL [Pourreza and Rafiei 2024] as the base model due to its strong performance in standard benchmarks like Spider. In simplified terms, it classifies the input query into one of three complexity levels—easy, medium, or hard—and then selects prompt examples from the corresponding category to support the generation of the final SQL query. To adapt it to our case study, we introduced the following modifications:

- **Translation of prompt examples to Portuguese:** Since our entire dataset is in Portuguese, all natural language examples used in the prompting process were translated accordingly.
- **Table filtering module before schema linking:** Given that the Censo Escolar tables contain hundreds of columns, we created a filtering mechanism to select only the relevant tables based on the user question. This step avoids exceeding the LLM token limitation.
- **Inclusion of categorical and temporal examples in the prompt:** To reflect the data characteristics, we added few-shot examples involving categorical values and time-based constraints (e.g., comparisons between years).
- **Evaluation module for performance analysis:** We implemented a custom evaluation pipeline to compute and compare predicted versus golden queries. This module enables both overall and case-specific analysis using the metrics described in Section 4.2.

The adapted prompts and the translation to English of the Portuguese questions shown in Section 5 can be found in the GitHub repository [2].

4.2. Performance Metrics

To evaluate the quality of the generated SQL queries, we adopted both standard and auxiliary metrics. The main metrics used were:

- **Execution Accuracy (EX):** Measures whether the predicted and golden queries return the same execution result.
- **Exact Set Match (EM):** Verifies if the predicted query matches the golden query exactly.

[2]https://github.com/froeska25/text-to-sql

- **Component Match (C-M):** Evaluates structural similarity by comparing individual SQL components (i.e., SELECT, FROM, JOIN, WHERE, GROUP BY, and subqueries).
- **Precision, Recall, and F1:** Derived from C-M to quantify partial correctness.
- **Cosine Similarity:** Used as an auxiliary metric to measure semantic similarity between the predicted and golden queries, based on vectorized SQL representations.

Given the temporal nature of our dataset, multiple valid SQL formulations may exist for the same question, especially when involving time-based filters, year-to-year comparisons, or historical aggregations. This reduces the reliability of string-based metrics such as Exact Match (EM) and Component Match (C-M), and highlights the need for semantic and more in-depth component-level evaluations.

During initial testing on real-world cases, it became evident that conventional metrics provided limited insight into model behavior. As a result, we prioritized a qualitative analysis to better understand the underlying challenges, as detailed in Section 5.

5. Case Study Results

Since conventional metrics failed to provide deep insight into model performance on our case set, we conducted a qualitative analysis of SQL outputs generated by our adapted model. We evaluated our set of natural language questions by comparing predicted SQL queries to manually crafted references. Key aspects considered include temporal reasoning, structural fidelity to the real-world schema and accurate representation of the intended meaning expressed in the Portuguese questions. Below, we present observations from ten selected examples, highlighting both successful reasoning patterns and common failure cases.

- **Q1:** Quais disciplinas são mais comuns nas turmas de ensino médio profissionalizante em escolas urbanas no último ano disponível?
  Predicted: `SELECT T1.disciplina_profissionalizante FROM Turma AS T1 JOIN Escola AS T2 ON T1.id_escola = T2.id_escola WHERE T1.etapa_ensino = 'ensino médio profissionalizante' AND T2.tipo_localizacao = 'urbana' AND T1.ano = (SELECT max(ano) FROM Turma)`
  Analysis: The model mistakenly treats a boolean field as a source of subject-level data, reflecting shallow schema understanding. However, it correctly performs the table join and applies appropriate filters. The golden query relies solely on the `Turma` table, so this variation would reduce Component Match despite logical similarities.

- **Q2:** Qual foi a média de idade dos docentes por estado ao longo dos últimos 5 anos de registro?
  Predicted: `SELECT AVG(idade), sigla_uf, ano FROM Docente WHERE ano BETWEEN 2017 AND 2021 GROUP BY sigla_uf, ano`
  Analysis: A structurally correct query with accurate aggregation and grouping. The only issue is the hardcoded year interval, which approximates but does not generalize to "last 5 years."

- **Q3:** Qual é o total de alunos que frequentam o ensino médio nas escolas públicas urbanas da Região Sul nos últimos 15 anos?
  Predicted: `SELECT COUNT(M.id) FROM Matricula M, Escola E WHERE M.id_escola = E.id_escola AND E.rede = 'publica'`

`AND E.tipo_localizacao = 'urbana' AND E.sigla_uf IN ('RS', 'SC', 'PR') AND M.ano BETWEEN (YEAR(CURRENT_DATE) - 15) AND YEAR(CURRENT_DATE) AND M.etapa_ensino = 'ensino médio'`
Analysis: The query uses appropriate filters and correctly identifies southern states and the JOIN logic. Moreover, the golden query uses a different table (`Turma`), which would penalize metric-based evaluation despite similar intent.

- **Q4:** Qual o número de matrículas por sexo em escolas rurais nos últimos 3 anos?
  Predicted: `SELECT count(T1.id_matricula), T1.sexo FROM Matricula AS T1 JOIN Escola AS T2 ON T1.id_escola = T2.id_escola WHERE T1.ano IN (2020, 2021, 2022) AND T2.tipo_localizacao = 'Rural' GROUP BY T1.sexo`
  Analysis: While the logic is sound and categorical filters are applied correctly, there is a hardcoded year interval, which approximates but does not generalize. The golden query used just table (`Escola`).
- **Q5:** Mostre a quantidade de matrículas de EJA (Educação de Jovens e Adultos) por estado nos últimos 10 anos.
  Predicted: `SELECT sigla_uf, COUNT(*) FROM Matricula WHERE eja = 1 AND ano BETWEEN 2012 AND 2022 GROUP BY sigla_uf`
  Analysis: The year range is imprecise and static, so temporal reasoning is not dynamically modeled, but overall logic is correct. Golden query uses just table `Escola`.
- **Q6:** Houve aumento no número de professores das redes públicas com grau de escolaridade de mestre se compararmos o ano atual e o anterior?
  Predicted: `SELECT (SELECT COUNT(*) FROM Docente WHERE mestrado = 1 AND ano = 2024) > (SELECT COUNT(*) FROM Docente WHERE mestrado = 1 AND ano = 2023)`
  Analysis: The model correctly identifies the comparison logic and relevant years but omits the filter for public institutions. This highlights how missing filters—while not invalidating execution—affect semantic equivalence. It also illustrates the ambiguity in how users phrase temporal comparisons.
- **Q7:** Quais as 10 escolas que apresentaram o maior crescimento percentual de matrículas em um único ano nos últimos 10 anos?
  Predicted: [No query returned]
  Analysis: The model likely failed, what could occur because of the complexity of self-joins and year-over-year comparisons, which require multi-step logic and analytical reasoning—capabilities not fully captured by most models.
- **Q8:** Qual o total de escolas fechadas no último ano?
  Predicted: `SELECT COUNT(*) FROM Escola WHERE ano = 2024 AND fechada = 'Sim'`
  Analysis: The model hallucinates a nonexistent column (`fechada`). The correct approach requires inferring closure by identifying schools present in year t but absent in year $t + 1$.
- **Q9:** Qual o total de alunos por raça matriculados nas redes de ensino infantil em 2007?
  Predicted: `SELECT sum(T1.quantidade_matricula_branca), sum(T1.quantidade_matricula_preta), sum(T1.quantidade_matricula_parda),`

`sum(T1.quantidade_matricula_amarela) FROM Matricula AS T1 JOIN Turma AS T2 ON T1.id_turma = T2.id_turma WHERE T1.ano = 2007 AND T2.etapa_ensino = 'ensino infantil'`

Analysis: The query structure is plausible, but it draws fields from the wrong tables. The actual race-related counts reside in `Escola`, not `Matricula` or `Turma`, as used.

- **Q10:** Qual o número total de escolas por região no ano de 2019?
  Predicted: `SELECT sigla_uf, COUNT(*) FROM Escola WHERE ano = 2019 GROUP BY sigla_uf`
  Analysis: Although syntactically valid, the query groups by state (`sigla_uf`) rather than region. The model fails to map states to broader regional categories, showing limitations in geographical abstraction.

Table 1 shows how traditional metrics—such as *Exact Match* and *Component Match*—fail to capture semantic inaccuracies or schema misuse. They underscore the need for evaluation methods that consider query intent, schema alignment, and dynamic temporal reasoning, particularly when working with complex, domain-specific datasets.

Table 1. Component Match (C-M) vs. Cosine Similarity Scores

Metric	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
C-M	0.0	0.1	N/A	0.7	0.2	0.1	0.0	0.5	0.6	0.8
Cosine Sim.	0.0	0.1	N/A	0.8	0.3	0.1	0.0	0.7	0.7	0.9

6. Conclusion and Future Work

This exploratory study highlighted how real-world, temporally grounded queries in Brazilian Portuguese pose unique challenges for current Text-to-SQL systems. Even when backed by strong base models such as DIN-SQL, we observed consistent limitations related to temporal reasoning, schema alignment, and the semantic ambiguity of natural language—particularly in multilingual settings.

Through a detailed analysis of representative queries, we found that traditional evaluation metrics often fail to reflect true semantic correctness or real-world utility. Hardcoded temporal values, hallucinated fields, omitted filters, and shallow understanding of schema structure were frequent issues. These findings underscore the need to move beyond accuracy-based evaluations and adopt frameworks that account for intent, context, and temporal fluidity.

As future work, we plan to (i) expand our manually annotated dataset to include a broader variety of temporal expressions and schema configurations, (ii) implement automatic evaluation methods that incorporate semantic or intent-based similarity, (iii) explore prompt tuning or retrieval-augmented techniques to dynamically guide LLMs toward better schema alignment and temporal abstraction, and (iv) evaluate other LLMs to assess potential gains in SQL generation. These efforts aim to advance the development of more robust, equitable, and explainable multilingual Text-to-SQL systems in public policy and education domains.

References

Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B., and Zhou, J. (2024). Text-to-SQL empowered by large language models: A benchmark evaluation. *Proc. VLDB Endow.*, 17(5):1132–1145.

Levene, M., Loizou, G., Levene, M., and Loizou, G. (1999). Temporal relational databases. *A Guided Tour of Relational Databases and Beyond*, pages 385–408.

Li, J., Hui, B., Qu, G., Yang, J., Li, B., Li, B., Wang, B., Qin, B., Geng, R., Huo, N., et al. (2023). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems*, 36:42330–42357.

Őzcan, F., Quamar, A., Sen, J., Lei, C., and Efthymiou, V. (2020). State of the art and open challenges in natural language interfaces to data. In *Proceedings of the 2020 ACM SIGMOD international conference on management of data*, pages 2629–2636.

Pourreza, M. and Rafiei, D. (2024). Din-SQL: Decomposed in-context learning of text-to-SQL with self-correction. *Advances in Neural Information Processing Systems*, 36.

Sun, R., Arik, S. Ö., Muzio, A., Miculicich, L., Gundabathula, S., Yin, P., Dai, H., Nakhost, H., Sinha, R., Wang, Z., et al. (2023). SQL-PaLM: Improved large language model adaptation for text-to-SQL (extended). *arXiv preprint arXiv:2306.00739*.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Visperas, M., Adoptante, A. J., Borjal, C. J., Abia, M. T., Catapang, J. K., and Peramo, E. (2023). On modern text-to-SQL semantic parsing methodologies for natural language interface to databases: A comparative study. In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 390–396. IEEE.

Wang, B., Shin, R., Liu, X., Polozov, O., and Richardson, M. (2019). Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. *arXiv preprint arXiv:1911.04942*.

Xu, X., Liu, C., and Song, D. (2017). Sqlnet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint arXiv:1711.04436*.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.