Addressing Data Scarcity: Augmentation Strategies in Text-to-SQL*

Felipe Aníbal Nunes Brito[1], **Kelly Rosa Braghetto**[1]

[1]Institute of Mathematics, Statistics and Computer Science – University of São Paulo
São Paulo — SP — Brazil

felipeanibal@usp.br, kellyrb@ime.usp.br

Abstract. *This work explores data augmentation as a strategy to overcome the scarcity of training data for Text-to-SQL translation models in low-resource domains. Six synthetic data generation strategies were applied to a novel dataset containing questions and SQL queries on Brazilian educational and geographic data. Experiments fine-tuning the LLM Mistral 7B demonstrated that using synthetic training data yields superior performance compared to using only original data. Furthermore, combining two augmented datasets allowed for even greater performance gains across several standard literature metrics, in addition to enabling the model to translate queries of greater structural complexity that were not handled by the models without augmented data.*

1. Introduction

The translation of natural language into SQL is a long-standing challenge in database research, dating back to early Natural Language Interfaces to Databases (NLIDB) [Androutsopoulos et al. 1995]. These initial approaches aimed to build systems capable of querying relational databases using English questions, relying primarily on rigid rules and grammars to map natural language to SQL syntax.

The advent of neural networks and large language models (LLMs) has revolutionized the field, introducing novel opportunities and enabling researchers to address these challenges with unprecedented efficacy [Li et al. 2023, Katsogiannis-Meimarakis and Koutrika 2023]. Nevertheless, training Text-to-SQL models requires vast amounts of annotated data, a requirement that can become prohibitively expensive or even intractable in specific contexts—particularly in low-resource languages, such as Portuguese, and in specialized domains, such as spatial querying. To address this bottleneck, this work proposes the application of data augmentation—a well-established strategy in other artificial intelligence domains—as a viable approach to mitigate data acquisition costs and improve the accuracy of Text-to-SQL models, using the Portuguese language and spatial queries as a case study.

We propose and evaluate a multi-faceted augmentation pipeline consisting of six distinct strategies, ranging from back-translation to LLM-driven paraphrasing and SQL structural variance. Our results demonstrate that these augmentation techniques can significantly improve Text-to-SQL model performance, increasing Exact Match accuracy

*This work was financed in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - Brazil (PIBIC Scholarship grant #2024-1955 / Universal grant #420623/2023-0) and by São Paulo Research Foundation (FAPESP) [grants #2023/00779-0 and #2023/18026-8].

from 5.2% to 17.2% in the best single-dataset configuration when fine-tuning the LLM Mistral 7B. Notably, combining multiple augmented datasets yielded the most robust results, achieving the highest Execution Accuracy of 48.3% and enabling the model to resolve complex 'hard' queries that had previously been handled unsuccessfully.

2. Theoretical Background

Formally, the *Text-to-SQL* task is defined as follows: given a natural language question regarding a relational database, the system must generate a syntactically valid SQL query that is semantically equivalent to the user's inquiry [Katsogiannis-Meimarakis and Koutrika 2023].

The challenges inherent to this task stem from two primary sources: the complexity of interpreting natural language and the structural rigidity of SQL [Katsogiannis-Meimarakis and Koutrika 2023]. While natural language is characterized by ambiguity, context-dependency, and noise, SQL requires strict syntactic precision. Questions such as 'What are the schools near the center of Salvador?' are intuitive for a human speaker but involve imprecise, context-dependent spatial concepts (e.g., the exact definitions of 'near' and 'center'). Conversely, queries that are conceptually straightforward in natural language—such as 'In each capital, which school has the highest number of other schools within a 1 km radius?'—can result in SQL statements that are excessively complex, computationally expensive, or even intractable to generate. While modern Machine Learning (ML) models effectively handle these contextual nuances, their success relies heavily on large-scale training data [Zhong et al. 2017].

2.1. Text-to-SQL Benchmarks

To address the scarcity of training data, academia introduced large-scale benchmarks that provide standardized frameworks for training and evaluating Text-to-SQL models. Notable examples include Spider [Yu et al. 2018], with over 10,000 question-query pairs across 200 databases, and BIRD [Li et al. 2023], comprising 12,741 pairs across 95 databases. Evaluating model performance on these benchmarks fundamentally relies on comparing the generated SQL against a known *ground-truth SQL*. Because this comparison can prioritize either strict syntax or semantic execution, robust assessment requires a combination of established metrics:

- **Exact Match (EM)**: This metric requires the generated SQL to be identical to the ground-truth SQL in terms of its exact character sequence and operators.
- **Component Matching or Structural Match (SM)**: Offering more flexibility, this metric evaluates whether the expected and generated queries share the same underlying SQL components (e.g., SELECT, JOIN, and WHERE clauses) rather than demanding an identical string match. While this mitigates penalties for superficial formatting differences, it fails to account for semantically equivalent queries that employ divergent structures (such as achieving the same result via JOINs versus nested subqueries).
- **Execution Accuracy (EX)**: This metric evaluates accuracy by executing both the generated SQL and the ground-truth SQL against the database. If the returned result sets are identical, the metric considers them equivalent. A notable limitation of execution accuracy is its susceptibility to false positives—instances in which an incorrect query coincidentally yields the correct result set on a specific database instance.

- **LLM as a Judge**: This metric employs an LLM to analyze the generated SQL against the ground-truth query and determine their semantic equivalence. Although this approach aims to overcome the rigidity of deterministic metrics, it remains susceptible to hallucination-induced false positives and lacks formal correctness guarantees.

2.2. Data Augmentation

According to [Li et al. 2021], the primary objective of data augmentation is to mitigate the distributional shift between the training data and the data encountered in real-world scenarios. In other words, augmentation is meant to inject novel variations into the training process, rendering the model more robust against edge cases and unforeseen conditions. To achieve this, data augmentation consists of applying specific transformations to samples from an existing dataset—introducing diversity while strictly preserving the core semantic characteristics and labels. In the context of the Text-to-SQL task, data augmentation entails generating novel pairs of natural language questions and corresponding SQL queries derived from the original dataset. As in machine translation, augmentation in the Text-to-SQL domain requires that any linguistic modifications to the natural-language question maintain a strict logical correlation with the underlying SQL query and vice-versa. This ensures that the generated synthetic pairs remain semantically valid and structurally consistent.

3. Related Work

Recent literature has demonstrated promising efforts to bridge the linguistic gap for the Portuguese language in Text-to-SQL tasks. For instance, the authors of *mRAT-SQL+GAP* [José and Cozman 2021] successfully fine-tuned transformer models for Portuguese through the translation of the English Spider dataset. Their methodology effectively validates a core premise of our research: that generating synthetic training data is a viable and necessary strategy for low-resource adaptation. However, their work focuses almost exclusively on direct translation—a powerful tool that can be significantly enhanced when coupled with other augmentation strategies.

Simultaneously, contemporary literature increasingly acknowledges the challenges that generalized models face when tackling niche domains. Spatial querying serves as a prime manifestation of this challenge. Recent work [Wang et al. 2025] has explicitly highlighted the vulnerability of standard models when confronted with PostGIS specific functions and spatial aggregations (e.g., `ST_Distance`, `ST_Contains`). While these studies effectively underscore the limitations of current architectures and introduce domain-specific datasets for evaluation, their strict reliance on high-resource languages (English and Chinese) exposes the ongoing, fundamental absence of training data for low-resource Text-to-SQL adaptation.

Highlighting the growing demand to fill these domain-specific low-resource gaps, the bilingual *text2SQL4PM* [Yamate et al. 2025] dataset was recently introduced for *process mining* in Portuguese. To actively generate such specialized data, frameworks like *Text2SQL-Flow* [Cai et al. 2025] and *SQL-PaLM* [Sun et al. 2024] have demonstrated that employing Large Language Models (LLMs) to systematically augment existing SQL queries is a highly effective direction. This validates yet another premise of our research: that curating specialized training data fundamentally increases model performance within

specific domains. Nevertheless, the manual curation of large-scale datasets from scratch remains costly in many contexts. Furthermore, while existing augmentation frameworks excel in English-centric environments, adapting these pipelines to low-resource languages introduces distinct linguistic and structural challenges.

Our work builds directly upon this foundation. Our goal is to demonstrate that data augmentation is a viable strategy that can be applied across highly specialized applications (such as geospatial queries) and low-resource domains (such as the Portuguese language). We propose a multi-faceted augmentation pipeline incorporating not only translation but also LLM-driven paraphrasing, cross-lingual synonym substitution, and structural SQL variance to efficiently create robust training datasets. Ultimately, the methodologies proposed herein provide a strategy for dataset augmentation, demonstrating how synthetic generation can improve Text-to-SQL performance in low-resource and domain-specific environments.

4. Data Augmentation

4.1. Baseline Dataset

As previously discussed, the manual curation of large-scale datasets for specialized domains is highly expensive. Therefore, we manually curated a compact dataset of 520 question-query pairs to serve as a seed for the augmentation strategies. This process involved drafting natural language questions alongside their corresponding SQL queries, followed by validating each query against a Database Management System (DBMS) to ensure syntactic correctness and data integrity.

These queries were formulated over a relational database schema comprising 15 tables constructed using microdata from the 2024 Basic Education Census (*Censo da Educação Básica*) and the School Catalog (*Catálogo de Escolas*), both provided by the Brazilian National Institute of Educational Studies and Research Anísio Teixeira (INEP). The Census data contains attributes of Brazilian schools, including enrollment numbers, available equipment, infrastructure conditions, administrative dependency (federal, state, municipal, or private), and specific locational characteristics (e.g., indigenous settlements or *quilombola* areas). Additionally, the School Catalog provides precise geographical coordinates and address data. Both datasets share a unique identifier for each educational institution, allowing their integration via relational joins.

Given the inherent spatial nature of the dataset, the PostGIS extension [PostGIS Steering Committee 2023] was employed to formulate a subset of the queries. By introducing geographic data types and advanced spatial functions—such as `ST_Distance` and `ST_Contains`—PostGIS enables the execution of complex geolocation queries that are not available in standard SQL.

To enable a granular analysis of model performance, the queries were categorized according to their **Complexity Level**. To quantify this, a scoring heuristic adapted from the Spider benchmark [Yu et al. 2018] was implemented, aggregating predefined weights to determine a final complexity score (S). Specifically, 1 point is added for each intermediate clause (e.g., `JOIN`, `GROUP BY`, `ORDER BY`), and 2 points are added for set operators, conditional expressions, or nested subqueries. Furthermore, to account for the inherent computational complexity of geospatial queries, the presence of any geometric

primitive or spatial function (e.g., `ST_Contains`, `POINT`, `POLYGON`) adds a 2-point penalty to the total score.

Using the accumulated score (S), queries were classified into four difficulty tiers: **Easy** ($S \leq 1$), involving single tables or simple filters; **Medium** ($1 < S \leq 3$), combining multiple common clauses or straightforward spatial queries; **Hard** ($3 < S \leq 6$), featuring subqueries, multiple joins, or spatial operations with complex aggregations; and **Extra Hard** ($S > 6$), exhibiting high nesting, set operators, or advanced geospatial logic.

4.2. Augmentation Strategies

To augment the baseline dataset, six distinct paraphrasing strategies were employed. Our methodology adopts two distinct approaches: (i) generating semantically equivalent Portuguese sentences for a fixed SQL query, and (ii) deriving alternative SQL syntaxes for a fixed Portuguese question.

The first category encompasses back-translation [Sennrich et al. 2016], multi-hop translation, synonym replacement [Wei and Zou 2019] in Portuguese, synonym replacement in English, and the generation of new Portuguese sentences via Large Language Models (LLMs) [Sun et al. 2024]. The underlying rationale for this approach is to expose the model to diverse linguistic formulations of identical intents. Consequently, the model learns that variations in tone, formality, and word order do not alter the target SQL representation. For instance, a formal directive such as 'List the schools located in the city of Salvador.' should map to the exact same SQL query as the more conversational inquiry, 'What are the names of the schools in Salvador?'.

A critical consideration in the broader context of data augmentation is the inherent trade-off between semantic fidelity (accuracy) and dataset variance (diversity) introduced by synthetic generation. According to [Feng et al. 2021], an ideal data augmentation strategy should maximize example variability without breaking the logical correspondence between the question and the answer. It is expected that the techniques discussed hereafter will introduce certain errors and inaccuracies during the generation of synthetic data. However, the anticipation is that the resulting gain in diversity will outweigh these imperfections. As a more objective metric, each technique will be evaluated based on the performance gains of the model trained on its generated data.

4.2.1. Back-translation

The back-translation technique for data augmentation [Sennrich et al. 2016] leverages the significant advancements made in machine translation over the past two decades. This method generates a paraphrase of a given text by translating it into an intermediate pivot language and subsequently returning it to the source language. The central premise is that, during the mapping process between distinct linguistic latent spaces, the original syntactic structure is altered while the underlying semantics are preserved [Mallinson et al. 2017].

A practical example illustrates this: the Portuguese question 'Quais ***colégios*** têm mais matrículas em Salvador?' can be translated into English as 'Which schools have the most enrollments in Salvador?'. However, upon translating it back to Portuguese, an acceptable output is 'Quais ***escolas*** têm mais matrículas em Salvador?'. Both

sentences preserve the exact same meaning, but the word 'colégios' is substituted with 'escolas'—both valid translations of the English term 'schools'. Thus, the back-translation technique effectively introduces lexical variety while maintaining semantic precision [Edunov et al. 2018].

4.2.2. Multi-hop Translation

Back-translation can be extended to incorporate multiple intermediate languages, a technique referred to as Multi-hop Translation. An example would be translating a sentence from Portuguese to English, from English to French, and finally from French back to Portuguese. Similar to single-hop back-translation, this approach leverages the morphological diversity across different languages. Consequently, word order may be modified, synonyms may be introduced, or the sentence may be entirely restructured to accommodate the linguistic disparities between languages. However, increasing the number of intermediate hops elevates the risk of accumulated semantic degradation.

To generate the augmented dataset utilizing these strategies, this study employed the Google Translator API in Python. Each sentence from the baseline dataset was translated into English and then back into Portuguese to execute the single-hop back-translation. To evaluate the Multi-hop Translation approach, sentences were translated from Portuguese to English, from English to French, and from French back to Portuguese. These strategies successfully yielded accurate paraphrases that injected structural diversity into the dataset. For instance, the query 'Liste o nome e o endereço de todos os *colégios* rurais no estado de Minas Gerais' was paraphrased as 'Enumere o nome e o endereço de todas as *escolas* rurais do estado de Minas Gerais'.

In certain instances, the translation pipeline failed to preserve the original semantic intent. For example, consider the sentence 'Liste as escolas no estado do Amazonas que possuem quadra de esportes coberta'. When translating the Portuguese word 'quadra' (sports court) to English, it became 'court'; this was subsequently translated into French as 'tribuna', and finally returned to Portuguese as 'tribunal' (legal court/tribunal). Consequently, the final sentence became 'Enumere as escolas do estado do Amazonas que cobriam o Tribunal de Esportes', where the expressions for 'covered sports court' were incorrectly translated as 'covered the Sports Tribunal'.

It is crucial to note that errors of this nature are highly dependent upon the specific machine translation implementations utilized. While certain translation tools may more accurately infer contextual nuances to provide precise translations, such errors remain an inherent risk. As discussed in the previous section, the expectation is that the localized semantic losses resulting from these translation errors will be offset by broader gains in lexical diversity — a trade-off that will ultimately be quantified by the performance metrics of models trained on the augmented data.

4.2.3. Synonym Substitution in Portuguese

Another approach to generating synthetic data is lexical paraphrasing via synonym substitution [Wei and Zou 2019]. This method aims to generate sentences with alternative

vocabulary while retaining the original semantic meaning, mitigating the risk of the Text-to-SQL model becoming overly reliant on specific terms present in the baseline dataset.

Effective synonym replacement requires identifying which words and expressions can represent the same concept within a given language and applying them correctly in the sentence. A traditional tool for this task is WordNet [Miller 1995], which organizes vocabulary based on semantic associations. Given the language requirements of this study, the Brazilian Portuguese adaptation of WordNet [de Paiva et al. 2012] was selected. Because WordNet does not store words in all their morphological variations, the synonyms found must be conjugated to agree with the rest of the sentence. To accomplish this, we employed spaCy [Honnibal et al. 2020], an NLP library capable of syntactic dependency parsing which allowed us to properly apply gender and number inflections to the newly inserted synonyms and their syntactic dependents.

However, correcting syntax alone is insufficient to produce coherent paraphrases. Synonymy is heavily context-dependent; substituting words without considering their context can result in semantic inconsistencies. In WordNet, for instance, the adjectives 'largest' and 'greatest' are treated as synonyms. In certain contexts, they are interchangeable, such as when referring to the 'greatest number of students' versus the 'largest number of students'. However, in a phrase like 'the greatest writer in English literature', substituting 'greatest' with 'largest' completely alters the meaning of the sentence. To resolve such contextual ambiguities, an evaluation strategy was implemented using BERTimbau [Souza et al. 2020], a Bidirectional Encoder Representations from Transformers (BERT) model pre-trained specifically for Portuguese. Inspired by [Salazar et al. 2020], the model was utilized to assess the naturalness of the generated phrases. This is achieved by masking the token associated with the substituted word and calculating the model's prediction loss for that token. A higher loss indicates that the word is less probable in that specific context. By generating multiple synonym candidates for a single sentence, we select the substitution that minimizes the model's loss.

While these strategies minimize errors, the generated sentences may still occasionally contain inconsistencies. These typically arise either from inaccuracies in the dictionary entries (e.g., using the archaism 'activa' as synonym for 'ativa'), from linguistic exceptions (like irregular plurals or verbs), the presence of slang, or failures in the syntactic dependency parsing.

4.2.4. Synonym Substitution in English

Considering that languages such as Portuguese possess significantly smaller volumes of computational linguistic resources compared to English, we explored a strategy of synonym substitution within a high-resource language [Magueresse et al. 2020], followed by translation. This strategy closely resembles the previously discussed method, with the distinction that the original sentence is first translated into English, synonym substitution is then applied, and the modified sentence is finally translated back into Portuguese. This approach benefits both from the superior maturity of English synonym databases and NLP libraries and from the added structural diversity inherent to the translation process—similar to the back-translation technique. This implementation utilized a standard BERT architecture, Google Translator API, the English WordNet, and the spaCy library.

4.2.5. Augmentation of Natural Language Questions using LLMs

A further approach to generating natural language paraphrases involves the use of Large Language Models (LLMs). This methodology stands out from traditional techniques due to the models' extensive pre-training contexts, which enable them to dynamically alter sentential attributes, such as tone and formality—aspects that are notoriously difficult to manipulate using the previously discussed methods [Dai et al. 2023].

In this study, we explored the application of the Gemini-2.5-flash model [Reid et al. 2024] to generate paraphrases from the baseline sentences. The primary advantage of this technique lies in its capacity to execute profound transformations on the textual surface while strictly preserving the underlying semantics. For example, the imperative command 'For each state, identify the municipality with the lowest proportion of schools with internet access' was reformulated as an interrogative: 'In each federative unit, which municipality presents the lowest percentage of schools provided with internet connection?'. In this instance, the sentence structure was changed from an imperative to a question, multiple expressions were replaced with synonyms, and the overall formality of the text was elevated. This is an example of how LLM-driven paraphrasing introduces substantial diversity without altering the original meaning of the query.

Furthermore, LLMs inherently encode what the literature defines as 'world knowledge', effectively functioning as implicit knowledge bases [Petroni et al. 2019]. A compelling demonstration of this capability is the query 'encontre os colégios militares localizados em Brasília' (find the military schools located in Brasília), which was paraphrased as 'liste os estabelecimentos de ensino militar situados na capital federal' (list the military educational establishments situated in the federal capital). This sentence exemplifies the model's ability to integrate external knowledge by accurately mapping 'Brasília' to 'capital federal'—a semantic deduction that would be unattainable using the strictly lexical or translational techniques discussed earlier.

4.2.6. Augmentation of SQL Queries Using LLMs

Analogous to the generation of natural language paraphrases, LLMs can be leveraged to synthesize structural variants of SQL queries. This methodology is corroborated by the work on SQL-PaLM [Sun et al. 2024], a state-of-the-art Text-to-SQL architecture. The capacity of large-scale models to generate structural paraphrases of identical SQL logic—such as substituting JOIN clauses with nested subqueries or employing Common Table Expressions (CTEs)—allows the translation model to be exposed to significantly greater syntactic diversity during training. Likewise, this work utilized the Gemini-2.5-flash model to ingest SQL queries and generate functionally equivalent alternatives. This process yielded a diverse corpus of SQL queries with varying degrees of complexity.

5. Evaluation

As discussed, the primary objective of data augmentation in Text-to-SQL tasks is to enhance model performance by expanding the training corpus. The underlying expectation is that introducing synthetic data will align the training data distribution more closely with the real-world data distribution. Consequently, models trained on augmented

datasets should exhibit greater generalization power and, ultimately, superior downstream performance. To empirically test this hypothesis and quantitatively compare the trade-offs of each augmentation strategy, 6 datasets were generated from each of the techniques described, and 12 were generated from combining two distinct strategies. Then, from each dataset, a machine learning model was trained (as shown in Figure 1). This allows for an evaluation of the impact of the synthetic data generated by each investigated technique.

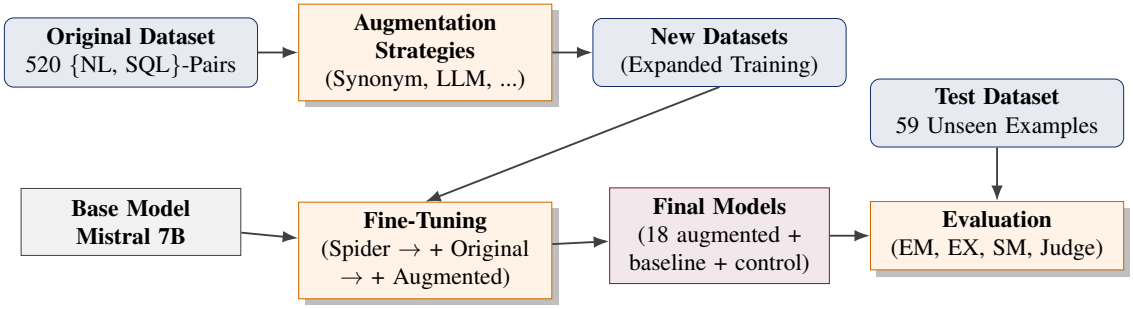

Figure 1. Overview of the methodology.

The training pipeline was structured into three phases, repeated for each dataset generated from the strategies previously described:

1. **Transfer Learning:** Initially, the model was fine-tuned using the Spider dataset [Yu et al. 2018]. Although this corpus consists of English-language data, this stage aims to converge the model upon the structural logic of SQL within multi-domain contexts, effectively leveraging Mistral's cross-lingual capabilities.
2. **Domain Adaptation (Baseline Setup):** The resulting model was subjected to a secondary fine-tuning phase utilizing the original, unaugmented dataset of 520 geo-referenced Portuguese-SQL query pairs. This model serves as the primary benchmark, representing the performance achieved without the introduction of synthetic data.
3. **Augmentation Evaluation:** Eighteen distinct model variations were subsequently trained, each incorporating different combinations of the synthetic data generated by the techniques detailed in Section 4.2.

Additionally, a control experiment was conducted wherein the model was trained on a duplicated version of the original dataset. This procedure is crucial to guarantee that any observed performance gains are directly attributable to the semantic diversity introduced by the augmented data, rather than simply resulting from prolonged exposure to the baseline data.

5.1. Text-to-SQL Model Tuning

This study aims to quantitatively evaluate data augmentation strategies for Text-to-SQL translation, advancing beyond simple qualitative assessments. To accurately isolate the impact of each synthetic data variant, our experimental design relies on a controlled baseline rather than pursuing absolute state-of-the-art (SOTA) performance or benchmarking against proprietary models. We selected the open-weight Mistral 7B model [Jiang et al. 2023] for this purpose, balancing cross-lingual capabilities with a compact architecture suitable for iterative fine-tuning. To optimize memory utilization, we employed Parameter-Efficient Fine-Tuning (PEFT) techniques—specifically LoRA

[Hu et al. 2021] and QLoRA [Dettmers et al. 2023]. All experiments were executed on a single NVIDIA A100 Tensor Core GPU (80GB VRAM). Under this hardware configuration, each training cycle was completed in approximately 30 minutes, providing a robust and efficient environment to measure the performance variations yielded by each augmentation technique.

5.2. Evaluation

Following the training phase, model performance was evaluated using an independent test set of 59 {question, SQL}-pairs. The compact size of this test set reflects the resource-intensive, manual curation required to build the underlying dataset. The complete corpus consists of 579 strictly curated pairs, manually written, reviewed and filtered to ensure high-quality ground truth. Because creating accurate SQL annotations is highly resource-intensive, we prioritized the training split to ensure the model received sufficient data volume for the augmentation experiments. While the test set is concise, it provides a strictly verified benchmark. For each natural language input, the fine-tuned Mistral 7B model generated a candidate query, which was subsequently evaluated with the metrics described in Section 2.1. For reproducibility purposes, all evaluation scripts, the prompts and the model utilized for the LLM judge, and the raw experimental results are available in the online repository (https://github.com/felipeanibal/da-text2sql-paper).

5.3. Results and Discussion

The performance of the trained Text-to-SQL models across each evaluation metric is shown in Table 1. The isolated application of individual data augmentation techniques yielded substantial gains. Under the optimal single-dataset configurations, Exact Match (EM) accuracy escalated from 5.2% to 17.2%, Structural Matching (SM) improved from 13.8% to 22.4%, Execution Accuracy (EX) increased from 32.8% to 39.7%, and the LLM-as-a-Judge metric surged from 39.7% to 53.4%. Composing multiple augmented datasets further boosted model performance, elevating the upper bounds across all evaluated metrics relative to the baseline. As shown in Table 1, EM reached 13.8%, SM reached 22.4%, EX peaked at 48.3%, and the LLM-as-a-Judge evaluation matched the single-dataset peak of 53.4%.

The highest Execution Accuracy (48.3%) was achieved by combining the Back-translation dataset with the LLM-driven NL paraphrasing dataset. Crucially, this specific combination successfully resolved 33.3% of queries classified as 'hard' under the EX metric—a threshold entirely unattained by the baseline architecture as shown in Table 2. Configurations combining LLM-driven NL variance with Portuguese synonym substitution, as well as LLM-driven SQL variance with English synonym substitution, also showed promise by resolving 16.7% of these complex queries.

While most strategies improved overall performance, certain configurations failed to surpass the baseline in deterministic and execution metrics. For instance, the Backtranslation + Synonym Substitution strategy underperformed across Exact Match (1.7% vs. 5.2%), Structural Match (8.6% vs. 13.8%), and Execution Accuracy (29.3% vs. 32.8%). However, these declines must be interpreted with caution, as strict metrics often penalize functionally correct queries that deviate structurally from the ground truth. For example, given the request *'Liste as escolas estaduais em Salvador'*, the ground truth selects only `nm_escola`, whereas the augmented models frequently retrieved

Table 1. Models' performance across each evaluation metric. In bold are the best results and underlined are the second best results for each metric.

Strategy	EM (%)	SM (%)	EX (%)	Judge
Baseline (Original Data)	5.2	13.8	32.8	39.7
Original Data Duplicated	3.4	6.9	27.6	36.2
Aug: Multi-hop	8.6	<u>19.0</u>	39.7	<u>51.7</u>
Aug: LLM Paraphrase (NL)	**17.2**	**22.4**	37.9	**53.4**
Aug: Backtranslation	5.2	12.1	32.8	43.1
Aug: Synonym Sub.	5.2	10.3	32.8	41.4
Aug: LLM Rewrite (SQL)	6.9	10.3	31.0	39.7
Aug: Trans + Synonym	8.6	10.3	29.3	48.3
Combo: Backtrans + LLM + NL	5.2	17.2	**48.3**	<u>51.7</u>
Combo: LLM + SQL + Multi	8.6	10.3	<u>41.4</u>	**53.4**
Combo: LLM + SQL + Trans + SYN	<u>13.8</u>	**22.4**	<u>41.4</u>	50.0
Combo: LLM + NL + Trans + SYN	12.1	17.2	<u>41.4</u>	**53.4**
Combo: Multi + Syn	10.3	13.8	36.2	46.6
Combo: LLM + NL + Syn	6.9	12.1	36.2	**53.4**
Combo: LLM + NL + Multi	6.9	17.2	34.5	**53.4**
Combo: Multi + Trans + SYN	8.6	13.8	34.5	**53.4**
Combo: Backtrans + Trans + SYN	8.6	13.8	34.5	48.3
Combo: LLM + NL + LLM + SQL	8.6	17.2	34.5	41.4
Combo: Backtrans + Syn	1.7	8.6	29.3	48.3
Combo: LLM + SQL + Syn	10.3	13.8	27.6	32.8

both `nm_escola` and `ds_endereco`. Although retrieving this extra column triggers a failure in both Exact Match and Execution Accuracy, the query remains valid and correctly answers the question. This interpretation is corroborated by the LLM-as-a-Judge evaluation, which focuses on semantic equivalence: the Backtranslation + Synonym Substitution model surpassed the baseline on the judge metric (48.3% vs. 39.7%). Conversely, the individual LLM Rewrite (SQL) and its combination with Synonym Substitution were the only augmentation strategies to perform at or below the baseline on the LLM-as-a-Judge metric, scoring 39.7% and 32.8%, respectively. We hypothesize that introducing profound structural SQL variations without a sufficiently large volume of training pairs hindered the model's ability to reliably map these complex syntactic variances, leading to genuine translation errors.

Finally, it is important to note that the control experiment explicitly showed that increasing the number of training epochs via data duplication does not yield better outcomes; instead, it results in a degradation in model performance.

6. Conclusion

This work demonstrates that data augmentation is an effective strategy for bridging linguistic and domain-specific gaps in Text-to-SQL tasks for low-resource languages, such as Portuguese. Our experimental results confirm this by showing that models trained on augmented datasets consistently outperform those trained solely on original data. Notably, the most significant performance gains are achieved by combining multiple

Table 2. Execution accuracy by difficulty level for all evaluated models.

Model	Easy (%)	Med. (%)	Hard (%)	Ex.Hard (%)
Baseline (Original Data)	43.2	25.0	0.0	**0.0**
Original Data Duplicated	32.4	33.3	0.0	**0.0**
Aug: Multi-hop	54.1	25.0	0.0	**0.0**
Aug: LLM Paraphrase (NL)	51.4	25.0	0.0	**0.0**
Aug: Backtranslation	40.5	33.3	0.0	**0.0**
Aug: Synonym Sub.	40.5	33.3	0.0	**0.0**
Aug: LLM Rewrite (SQL)	40.5	25.0	0.0	**0.0**
Aug: Trans + Synonym	37.8	25.0	0.0	**0.0**
Combo: Backtrans + LLM + NL	**59.5**	**50.0**	0.0	**0.0**
Combo: LLM + SQL + Multi	43.2	**50.0**	**33.3**	**0.0**
Combo: LLM + SQL + Trans + SYN	48.6	41.7	16.7	**0.0**
Combo: LLM + NL + Trans + SYN	54.1	33.3	0.0	**0.0**
Combo: Multi + Syn	43.2	41.7	0.0	**0.0**
Combo: LLM + NL + Syn	48.6	16.7	16.7	**0.0**
Combo: LLM + NL + Multi	45.9	25.0	0.0	**0.0**
Combo: Multi + Trans + SYN	37.8	**50.0**	0.0	**0.0**
Combo: Backtrans + Trans + SYN	45.9	25.0	0.0	**0.0**
Combo: LLM + NL + LLM + SQL	48.6	16.7	0.0	**0.0**
Combo: Backtrans + Syn	32.4	41.7	0.0	**0.0**
Combo: LLM + SQL + Syn	37.8	16.7	0.0	**0.0**

augmentation techniques, rather than relying on a single method.

The efficacy of Natural Language diversification strategies is largely attributable to the injection of lexical and syntactic variance. The combination of LLM-driven structural variations and translational artifacts successfully conditioned the model to remain resilient against the diversity of natural Portuguese user inputs. This resilience was particularly evident in the combined dataset experiments, which enabled the model to resolve more complex queries that the baseline architecture could not process.

As future work, we aim to generalize these findings by applying our augmentation pipeline to additional specialized domains and testing its performance on larger datasets.

Generative AI Disclosure

The authors disclose the use of Generative AI tools (Large Language Models) to assist in the linguistic revision, structural formatting, and rephrasing during the drafting of this paper, as well as in the generation of source code used for the experimental evaluation. The authors assume full and integral responsibility for the technical content, architectural proposals, experimental results, and code correctness.

References

Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. (1995). Natural language interfaces to databases-an introduction. *Natural language engineering*, 1(1):29–81.

Cai, Q., Liang, H., Xu, C., Xie, T., Zhang, W., and Cui, B. (2025). Text2SQL-flow: A robust SQL-aware data augmentation framework for text-to-SQL. *arXiv preprint arXiv:2511.10192*.

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., et al. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *IEEE Transactions on Big Data*.

de Paiva, V., Rademaker, A., and de Melo, G. (2012). OpenWordNet-PT: An open brazilian wordnet for reasoning. In *Proceedings of the 6th Global Wordnet Conference (GWC 2012)*, pages 353–360. The Global Wordnet Association.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Edunov, S., Ott, M., Auli, M., and Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500. ACL.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for natural language processing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. `https://spacy.io/`.

Hu, E. J., Shen, Y., Wallis, P., et al. (2021). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., et al. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*.

José, M. A. and Cozman, F. G. (2021). *mRAT-SQL+GAP: A Portuguese Text-to-SQL Transformer*, page 511–525. Springer International Publishing.

Katsogiannis-Meimarakis, G. and Koutrika, G. (2023). A survey on deep learning approaches for text-to-SQL. *The VLDB Journal*, 32(4):905–936.

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.

Li, Y., Hu, Y., et al. (2021). Data augmentation for text-to-SQL. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Magueresse, A., Vincent, C., and Meignier, S. (2020). Low-resource languages: a review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Mallinson, J., Sennrich, R., and Lapata, M. (2017). Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 881–893.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Petroni, F., Rocktäschel, T., Lewis, P., et al. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

PostGIS Steering Committee (2023). PostGIS 3.4.0 Developer Guide. `https://postgis.net/documentation/`.

Reid, M., Savinov, N., Teplyashin, D., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712. ACL.

Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: Pretrained BERT models for brazilian portuguese. In *9th Brazilian Conference on Intelligent Systems (BRACIS)*.

Sun, R., Arik, S. Ö., Nakhost, H., Dai, H., Sinha, R., Yin, P., and Pfister, T. (2024). SQL-PaLM: Improved large language model adaptation for text-to-sql. *Transactions on Machine Learning Research*.

Wang, H., Guo, L., Liang, Y., Liu, L., and Huang, J. (2025). GPT-Based text-to-SQL for spatial databases. *ISPRS International Journal of Geo-Information*, 14(8).

Wei, J. and Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Yamate, B. Y., Neubauer, T. R., Fantinato, M., and Peres, S. M. (2025). Text-to-SQL oriented to the process mining domain: A PT-EN dataset for query translation. *arXiv preprint arXiv:2509.09684*.

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 hierarchical dataset for complex semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3911–3921.

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