The Impact of Language and Portuguese Intralinguistic Variation on NL2SQL

Ricardo Trainotti Rabonato , Lilian Berton

[1]Universidade Federal de São Paulo (UNIFESP)
São José dos Campos – SP – Brazil

trainotti.ricardo@unifesp.br, lberton@unifesp.br

Abstract. *Natural Language to Structured Query Language (NL2SQL) systems have advanced with large language models (LLMs), yet evaluations remain centered on English, leaving open how these systems behave across languages and language varieties. We aim to contribute to this by analyzing two state-of-the-art LLMs (LLaMA 3.3 70B and Qwen 3.6 35B, both zero-shot) on the WikiSQL benchmark in English, Brazilian Portuguese (PT-BR), and European Portuguese (PT-PT), keeping database schemas and inference configuration fixed. Translations were validated through distance metrics (cosine similarity 0.966), contrastive linguistic analysis confirming varietal features (e.g., cleft constructions 36.8× more frequent in PT-PT), and multi-model judgment (85% / 81% semantic equivalence, 95.5% PT-BR and 96.0% PT-PT inter-judge agreement). Results show a marked drop from English (47.25% EX) to Portuguese (PT-BR: 40.03%; PT-PT: 39.34%), with the PT-BR vs. PT-PT difference reaching statistical significance ($p = 0.019$). English-only successes outnumber Portuguese-only successes nearly 10:1. A replication on a second model (Qwen 3.6 35B) confirms the English-Portuguese gap but suggests that intralinguistic effects may be model-dependent. Stratified analyses show semantic complexity and quantification amplify degradation, particularly for COUNT queries. These findings point to persistent English bias in NL2SQL and measurable effects of intralinguistic variation, even between closely related varieties.*

1. Introduction

Language models have played an increasingly important role in *semantic parsing*, particularly in the conversion of natural language into SQL queries (NL2SQL) [Liu et al. 2024, Zhao et al. 2025]. Large language models (LLM) now enable complex query generation from natural language prompts in *zero-shot* or *few-shot* settings, reducing reliance on specialized architectures and expanding NL2SQL applicability in analytical and corporate systems [Katsogiannis-Meimarakis and Koutrika 2023, Fan et al. 2024]. Despite these advances, research remains strongly centered on English-language data and benchmarks such as WikiSQL and Spider [Zhong et al. 2017, Yu et al. 2018], leaving open how NL2SQL systems behave in other languages—especially those with substantial internal variation, like Portuguese.

Portuguese exhibits a well-established division between its two most widely used written varieties: Brazilian Portuguese (PT-BR) and European Portuguese (PT-PT). Although mutually intelligible, these varieties display differences in vocabulary, syntax, polysemy, lexical preferences, and register [Azevedo 2005]. Prior work indicates that linguistic differences between varieties of the same language can affect multilingual model performance, particularly when these varieties are unevenly represented in pretraining

data [Pires et al. 2019]. However, evidence remains scarce on how such differences manifest in tasks with high semantic complexity—such as SQL query generation, which demands precise intent interpretation, entity identification, accurate column mapping, and multi-step logical reasoning [Liu et al. 2024, Zhao et al. 2025].

This gap matters because current LLMs exhibit a strong bias toward English, which dominates both pretraining corpora and evaluation benchmarks. When deploying NL2SQL for Portuguese, two distinct linguistic factors come into play: (i) cross-lingual transfer from English to Portuguese [Shi et al. 2022], and (ii) intralinguistic variation between PT-BR and PT-PT [Preda et al. 2024]. While cross-lingual transfer has received considerable attention, the impact of intralinguistic variation remains largely unexplored—despite its practical relevance for applications serving users across Brazil and Portugal, where the same system must handle both varieties.

In this work, we investigate how general-purpose LLMs perform NL2SQL when exposed to questions in English, PT-BR, and PT-PT. Our primary experiments use LLaMA 3.3 70B on the full WikiSQL benchmark [Zhong et al. 2017], with a replication on Qwen 3.6 35B to assess cross-model generalization. Questions were translated into both Portuguese varieties and validated through a multi-layered protocol combining distance metrics, contrastive linguistic analysis, and multi-model judgment. By keeping the database schema, instructions, and hyperparameters identical across languages, we isolate the impact of input language and intralinguistic variation. We evaluate performance using *Execution Accuracy* and *Logical Form Accuracy*, stratified by query complexity and aggregation type, complemented by qualitative error analysis.

The contributions of this work are threefold:

- A trilingual evaluation (EN, PT-BR, and PT-PT) of WikiSQL using a controlled NL2SQL setup with two models from different families, quantifying the impact of input language and assessing cross-model consistency.
- The first analysis of Portuguese intralinguistic variation on NL2SQL performance, identifying error patterns tied to semantic complexity and aggregation type.
- A parallel corpus of WikiSQL questions translated into PT-BR and PT-PT, that will be released for future research on NL2SQL in Portuguese.

2. Related Work

NL2SQL has evolved from early sequence-to-sequence architectures such as Seq2SQL [Zhong et al. 2017] and SQLNet [Xu et al. 2017] to specialized models with structure-aware decoding, including RAT-SQL [Wang et al. 2020] and PICARD [Scholak et al. 2021]. Standardized benchmarks like WikiSQL and Spider [Yu et al. 2018] enabled systematic comparisons across models, yet the field remains strongly centered on English.

With large language models, NL2SQL can be performed via few-shot prompting or instruction tuning, without task-specific architectures [Katsogiannis-Meimarakis and Koutrika 2023, Fan et al. 2024]. Models such as T5 [Raffel et al. 2020], Codex [Chen et al. 2021], and LLaMA [Touvron et al. 2023] generate high-quality SQL queries even for unseen databases—but they are predominantly trained on English corpora, with limited exposure to technical Portuguese.

Portuguese NLP resources exist—BERTimbau [Souza et al. 2020], Universal Dependencies treebanks [Rademaker et al. 2017], BrWaC [Wagner Filho et al. 2018]—yet focus predominantly on Brazilian Portuguese. Cross-lingual transfer studies show that multilingual models such as mBERT degrade on underrepresented language variants [Pires et al. 2019], though these analyses have been restricted to lower-level tasks like POS tagging, leaving open the question for semantically complex tasks.

Recent work has begun addressing multilingual NL2SQL [Shi et al. 2022] and Portuguese dialect identification [Preda et al. 2024]. To our knowledge, however, no prior study systematically investigates the impact of Portuguese intralinguistic variation on NL2SQL, nor provides parallel benchmarks covering PT-BR and PT-PT. This gap directly motivates the present work.

3. Method

Our experimental design isolates the impact of input language and intralinguistic variation by varying only the language of the natural language question, while holding all other factors constant: database schema, gold SQL queries, and inference parameters.

3.1. Base Benchmark

We use WikiSQL, one of the most widely adopted datasets for evaluating NL2SQL systems [Zhong et al. 2017]. WikiSQL pairs natural language questions with corresponding SQL queries over real Wikipedia tables. Each example includes a question, a structured SQL representation, and a target table, enabling both execution-based and structural evaluation. Although originally proposed in English, WikiSQL's relatively simple structure—selection, aggregation, and filtering conditions—makes it well suited for controlled comparisons across linguistic variants.

3.2. Question Translation

Starting from the original English questions in the WikiSQL development set, we constructed two translated versions: PT-BR and PT-PT. Translations were produced automatically using the same LLaMA 3.3 70B model used in the main experiment, with explicit instructions to avoid simplification and to maintain interrogative structures and lexical choices compatible with each variety. No manual post-editing was performed, nor were cultural adaptations or free reformulations introduced. Differences between PT-BR and PT-PT thus predominantly reflect linguistic preferences learned during pretraining. The goal was to produce linguistically plausible, semantically equivalent versions of the original questions, without task-specific optimization for NL2SQL.

3.2.1. Translation Validation

We assessed translation quality through three complementary layers: automatic distance metrics over the full corpus, contrastive linguistic analysis, and multi-model judgment on a stratified sample.

Distance Metrics. Comparing each PT-BR/PT-PT pair across the full corpus (N=8,421) revealed high semantic equivalence with consistent surface divergence. Mean cosine similarity computed with multilingual sentence embeddings[1] was 0.966 (SD=0.058), confirming near-perfect semantic preservation. BLEU-4 averaged 0.709 (SD=0.298) and normalized Levenshtein distance 0.104 (SD=0.113)—indicating systematic lexical and syntactic differences between variants despite shared meaning. Token count ratios averaged 0.986 (SD=0.157), ruling out systematic lengthening or shortening in either variant.

Linguistic Characterization. To verify that translations reflect authentic varietal differences rather than superficial lexical substitution, we analyzed syntactic features known to distinguish the two varieties. Table 1 summarizes the results.

Feature	PT-BR	PT-PT
Cleft constructions ("é que")	4	147
Gerund periphrases ("está correndo")	30	0
A + infinitive ("está a correr")	0	11
Enclitic pronoun placement	0	24
Proclitic pronoun placement	719	659
Article + possessive ("o meu")	7	50

Table 1. Frequency of varietal linguistic features in the PT-BR and PT-PT translations (N=8,421). The sharp contrasts (e.g., cleft constructions 36.8× more frequent in PT-PT, gerund periphrases exclusive to PT-BR) validate that translations authentically reflect each variety's syntactic preferences.

Cleft constructions—a well-known marker of European Portuguese—appear 36.8 times more frequently in PT-PT than in PT-BR. Gerund periphrases occur exclusively in PT-BR, while the a + infinitive construction occurs exclusively in PT-PT, mirroring the documented aspectual divide between the varieties. Enclisis appears only in PT-PT translations, and the rate of article + possessive sequences is over seven times higher in PT-PT (50 vs. 7). These patterns confirm that the translations respect varietal syntactic norms.

Multi-Model Quality Judgment. We further validated a stratified sample of 200 question pairs—100 where English succeeded but Portuguese failed, 50 where all languages succeeded, and 50 random—using three independent LLMs as judges (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B). Each judge assessed translations along three criteria using a standardized scale: *semantic equivalence* (YES / PARTIAL / NO), *varietal naturalness* (NATURAL / ACCEPTABLE / UNNATURAL), and *complexity preservation* (YES / PARTIAL / NO), with judges instructed to respond exclusively in JSON format. Inter-judge agreement was high: at least two of three judges concurred in 95.5% of cases for PT-BR and 96.0% for PT-PT.

The majority vote was "YES" for semantic equivalence in 85.0% of PT-BR and 81.0% of PT-PT translations, with only 6.0% and 5.5% rated "NO." Complexity preservation was near-perfect (97.0% PT-BR, 95.0% PT-PT). Varietal naturalness ratings were "NATURAL" or "ACCEPTABLE" for 67.5% of PT-BR and 72.0% of PT-PT translations.

[1] Using `paraphrase-multilingual-MiniLM-L12-v2` from the Sentence Transformers library.

Taken together, these analyses indicate that the translations are semantically faithful, exhibit authentic varietal features, and are free of systematic artifacts that could confound the NL2SQL evaluation.

3.3. Model Configuration

We use LLaMA 3.3 70B [Meta AI 2024] in a *zero-shot* setting across all conditions, without fine-tuning or task-specific training. This lets us analyze how a widely used generalist model responds to linguistic variation in the input. We stress that our goal is not to assess absolute NL2SQL quality, but to measure relative performance differences arising exclusively from the language of the question.

To ensure direct comparability, we keep the following fixed across all experiments:

- table schemas (in English, as in the original WikiSQL);
- input format;
- instruction prompt (in English, unchanged across conditions);
- decoding and inference parameters;
- number of attempts per query.

Under this design, any observed variation in performance is attributable exclusively to the language of the natural language question.

For cross-model validation, we replicate the main experiment on a stratified sample of 500 examples using Qwen 3.6 35B [Qwen Team 2026], a Mixture-of-Experts model from a different family (Alibaba). All other configuration parameters (schemas, prompts, translations, and decoding settings) remain identical.

3.4. Evaluation Protocol

Evaluation was conducted separately for each linguistic condition (EN, PT-BR, PT-PT) on the same examples and WikiSQL tables. We employ two complementary metrics:

- **Execution Accuracy (EX)**: proportion of queries whose execution result matches the gold result when both queries are run on the database, regardless of surface differences in SQL formulation. Two queries are considered equivalent if they return identical result sets, even if their SQL structures differ;
- **Logical Form Accuracy (LF)**: proportion of queries whose parsed representation exactly matches the gold query's parsed representation. Following the WikiSQL parser [Zhong et al. 2017], we convert each SQL dictionary into a `Query` object and consider a prediction correct if its parsed structure is identical to the gold structure (qp == qg), including all conditions, aggregations, and ordering.

Execution Accuracy captures functional behavior in practical scenarios, while Logical Form Accuracy provides a stricter measure of structural correspondence. Both were computed identically across conditions.

3.5. Analysis by Query Type

We stratify results by structural characteristics of the gold SQL queries: number of conditions in the `WHERE` clause, and aggregation type in the `SELECT` clause (`COUNT`, `SUM`, `AVG`, etc.). This lets us examine whether linguistic effects are uniform or intensify with query complexity.

3.6. Qualitative Error Analysis

We complement quantitative results with exploratory inspection of selected examples, focusing on cases where the English query succeeds but Portuguese versions fail, and where PT-BR and PT-PT diverge. The analysis is not exhaustive; its purpose is to illustrate recurring patterns—difficulties with quantification, multi-condition combinations, or interrogative constructions—that support the statistical trends identified in our experiments.

4. Results

Evaluation was conducted on the WikiSQL development set (8,421 examples per condition). Table 2 summarizes overall performance.

Language	EX (%)	LF (%)
English (EN)	**47.25**	28.93
Brazilian Portuguese (PT-BR)	**40.03**	22.53
European Portuguese (PT-PT)	**39.34**	22.30

Table 2. Execution (EX) and Logical Form (LF) accuracy of LLaMA 3.3 70B (zero-shot) on the WikiSQL development set, with schemas and inference parameters held constant across English, PT-BR, and PT-PT.

4.1. Overall Performance

Performance drops sharply from English to Portuguese. The model reaches 47.25% Execution Accuracy in English, falling to 40.03% for PT-BR and 39.34% for PT-PT—a gap of 7–8 percentage points attributable entirely to input language, since schemas and SQL targets remained fixed. Logical Form Accuracy follows the same pattern (28.93% EN vs. 22.53% PT-BR and 22.30% PT-PT).

4.2. Comparison between PT-BR and PT-PT

Differences between Portuguese variants are smaller but consistent: PT-BR outperforms PT-PT by 0.69 points in EX and 0.23 in LF, persisting across both evaluation dimensions.

4.3. Statistical Tests

We applied the paired McNemar test to all language pairs. Differences are highly significant for EN vs. PT-BR ($\chi^2 = 289.2$, $p < 0.001$) and EN vs. PT-PT ($\chi^2 = 341.2$, $p < 0.001$). The PT-BR vs. PT-PT difference also reaches significance ($\chi^2 = 5.51$, $p = 0.019$): intralinguistic variation has a measurable, non-random effect. Discordant pairs between Portuguese variants account for 7.0% of examples (590 of 8,421), with a slight PT-BR advantage (324 vs. 266 exclusive correct predictions).

4.4. Error Overlap Analysis

Table 3 shows how correctness patterns distribute across languages.

English-only successes (8.9%) outnumber Portuguese-only successes (1.1% and 0.9%) by nearly an order of magnitude, reinforcing that input language—not random variation—drives the gap. Among Portuguese errors, both variants share the same outcome in the vast majority of cases (4,784 both incorrect vs. 590 discordant), consistent with a common English-centric pretraining bias. The 590 discordant cases are where intralinguistic variation leaves a measurable footprint.

Pattern	Count	%
All three correct	2,806	33.3
All three incorrect	4,035	47.9
EN only correct	749	8.9
PT-BR only correct	92	1.1
PT-PT only correct	74	0.9
EN and PT-BR correct (PT-PT wrong)	232	2.8
EN and PT-PT correct (PT-BR wrong)	192	2.3
PT-BR and PT-PT correct (EN wrong)	241	2.9

Table 3. Overlap of execution correctness across English (EN), Brazilian Portuguese (PT-BR), and European Portuguese (PT-PT) for the 8,421 queries. "EN only correct" cases outnumber Portuguese-only successes nearly 10:1, indicating strong English-centric bias.

4.5. Performance by Query Type and Aggregation

Tables 4 and 5 stratify Execution Accuracy by WHERE conditions and aggregation operator.

Query Type	N	EN	PT-BR	PT-PT
Conds = 0	64	**78.12**	73.44	67.19
Conds = 1	5,755	**53.19**	45.66	45.54
Conds = 2	2,082	**33.77**	26.95	24.93
Conds ≥ 3	520	**31.73**	25.96	25.00

Table 4. Execution accuracy stratified by number of conditions in the `WHERE` clause (N = number of examples per class). The English-Portuguese gap widens with query complexity, showing that semantic complexity amplifies linguistic mismatch.

Performance degrades as complexity increases; while all languages suffer, the English-Portuguese gap remains evident throughout.

Aggregation	N	EN	PT-BR	PT-PT
NONE	6,017	49.04	42.94	42.70
COUNT	779	29.91	14.76	11.17
MAX	507	50.89	43.98	41.62
MIN	468	52.99	44.23	42.09
SUM	321	41.12	32.71	35.51
AVG	329	47.72	41.64	41.03

Table 5. Execution accuracy by aggregation operator in the `SELECT` clause (N = number of examples per operator). The `COUNT` operator shows the most severe degradation in Portuguese, with PT-PT accuracy (11.17%) reaching only one-third of the English baseline (29.91%).

The `COUNT` operator shows the most severe degradation: PT-PT accuracy (11.17%) is roughly one-third of EN accuracy (29.91%). Quantification operators (`SUM`, `AVG`) also amplify sensitivity to input language, particularly in European Portuguese.

4.6. Cross-Model Replication

To assess whether the observed patterns are specific to LLaMA 3.3 70B, we replicated the experiment on a stratified random sample of 500 examples using Qwen 3.6 35B

[Qwen Team 2026], a model from a different family and architecture (Mixture-of-Experts, Alibaba). We kept the same translations, schemas, prompts, and inference configuration.

Table 6 summarizes the results. Both models exhibit a clear performance drop from English to Portuguese, confirming that the English bias is not an artifact of a single model. The gap is larger for LLaMA (6.20 pp) than for Qwen (3.00 pp), but the direction is consistent.

Model	EN	PT-BR	PT-PT	Δ EN$-$PT
LLaMA 3.3 70B	44.80	40.40	38.60	6.20
Qwen 3.6 35B	51.60	48.20	48.60	3.00

Table 6. Cross-model comparison of Execution Accuracy (EX) on a stratified random sample of 500 examples. LLaMA 3.3 70B and Qwen 3.6 35B both show a consistent performance drop from English to Portuguese, confirming that the English-centric bias is not model-specific. Δ EN$-$PT denotes the accuracy gap between English and the average of PT-BR and PT-PT.

We note that LLaMA's accuracy on this 500-example sample (44.80% EN) differs slightly from the full corpus (47.25%) due to the stratified sampling, which oversampled complex queries relative to their corpus proportion.

We applied the paired McNemar test to Qwen's predictions on the 500-example sample. None of the pairwise comparisons reached statistical significance at the 5% level: EN vs. PT-BR ($\chi^2 = 3.24$, $p = 0.072$), EN vs. PT-PT ($\chi^2 = 2.42$, $p = 0.120$), and PT-BR vs. PT-PT ($\chi^2 = 0.03$, $p = 0.864$). We attribute this primarily to the reduced sample size (500 vs. 8,421 for LLaMA), which limits statistical power. The numerical trends, however, align with LLaMA's: English outperforms both Portuguese varieties, and PT-BR and PT-PT are virtually indistinguishable in Qwen's predictions (48.20% vs. 48.60%).

The intralinguistic pattern differs between models. LLaMA consistently favors PT-BR over PT-PT across all query types and aggregation operators, with a statistically significant gap on the full corpus. Qwen shows no systematic advantage for either variant: the direction varies across strata and the gap is negligible (0.4 pp). This suggests that while the English-Portuguese gap is a robust cross-model phenomenon, the subtler effects of intralinguistic variation may depend on model-specific factors such as pretraining data composition and tokenization.

For the COUNT operator, LLaMA shows the most severe degradation in Portuguese (25.0% EN vs. 8.3% for both PT-BR and PT-PT on the sample), while Qwen performs comparably across languages (50.0% EN vs. 54.2% PT-BR vs. 50.0% PT-PT). We caution that the COUNT stratum contains only 48 examples, so these figures should be interpreted as indicative rather than conclusive.

4.7. Summary

Across all analyses, input language emerges as a decisive factor in NL2SQL performance for both models tested. The English-Portuguese gap is consistent in direction and substantial in magnitude: 7–8 percentage points for LLaMA on the full corpus and 3–6 points for both models on the stratified sample, with semantic complexity and quantification operators amplifying the degradation. The intralinguistic pattern, however, proved

model-dependent. LLaMA shows a statistically significant but small PT-BR advantage ($p = 0.019$), while Qwen exhibits no systematic preference for either variant. This suggests that while English-centric bias is a robust cross-model phenomenon, the subtler effects of Portuguese intralinguistic variation may depend on specific pretraining data composition and tokenization strategies, warranting further investigation across a broader range of models.

5. Discussion

5.1. English Bias in NL2SQL

The sharp drop from English to Portuguese—under identical schemas and SQL targets—points to a pronounced English bias in these models, likely rooted in the predominance of English in pretraining data and in NL2SQL benchmarks. Since database schemas remained in English, the degradation reflects the added difficulty of mapping Portuguese questions to English schema elements: the core cross-lingual challenge. Stratified results show this is not uniform: queries with multiple conditions and quantification operators amplify the degradation, indicating that semantic complexity compounds linguistic mismatch.

5.2. Intralinguistic Variation: PT-BR vs. PT-PT

The statistically significant PT-BR vs. PT-PT difference in Llama ($p = 0.019$), calls for explanation, though we note that this pattern did not replicate in Qwen, where the two variants performed nearly identically (48.20% vs. 48.60%). For LLamA, we see two plausible factors. First, the imbalance in pretraining data representation, where PT-BR tends to be more prevalent, may give the model marginally better adaptation to its lexical and syntactic patterns. Our linguistic validation confirms that translations exhibit authentic varietal features: cleft constructions 36.8$\times$ more frequent in PT-PT, enclisis exclusive to PT-PT, gerund periphrases exclusive to PT-BR. Llama thus faces genuinely different input distributions across variants.

Second, automatic translation itself may introduce subtle artifacts. PT-BR translations received slightly higher semantic equivalence ratings (85.0% vs. 81.0%), which could contribute to the gap. We treat these as explanatory hypotheses rather than definitive conclusions, given the absence of human-translated references. The fact that Qwen does not exhibit a PT-BR advantage suggests that intralinguistic effects, unlike the English-Portuguese gap, are not uniform across models and may depend on specific pretraining data mixtures or tokenization choices.

5.3. Qualitative Error Patterns

Inspecting discordant cases reveals patterns beyond lexical differences. Identical questions across variants sometimes produced different errors—schema linking failures independent of translation quality. Semantically equivalent quantification expressions (*"pontuação total"* vs. *"total de pontos"*) triggered different aggregation operators, suggesting sensitivity to how totality is syntactically framed. Temporal expressions with opposite interpretations (*"primeiros anos"* vs. *"ano mais antigo"*) led to opposite operators (MIN vs. MAX). These cases show how minor intralinguistic differences propagate

into structurally distinct SQL queries, aligning with broader findings in semantic parsing where subtle linguistic variation produces systematic errors in column and operator identification.

The severe degradation of `COUNT` (29.91% EN vs. 14.76% PT-BR vs. 11.17% PT-PT) reflects a structural challenge absent from other operators. Whereas `MAX`, `MIN`, `SUM`, and `AVG` typically anchor to an explicit numeric column, `COUNT` must be inferred from the question's semantic intent (e.g., *"quantos jogadores"* vs. *"qual o número de"* vs. *"conte"*), offering no direct lexical anchor to guide schema linking. This makes the operator uniquely dependent on cross-lingual parsing of quantification intent—precisely where the model's English-centric pretraining bias is most consequential. The pattern aligns with the sensitivity to syntactic framing observed across Portuguese variants, where semantically equivalent expressions triggered different aggregation operators.

5.4. Implications

These findings underscore the need for multilingual and multi-varietal evaluation in NL2SQL. Relying solely on English benchmarks risks overestimating system robustness. For practitioners, even small differences between language varieties can produce measurable degradation, which is a concern in business intelligence and data analysis applications serving diverse Lusophone users. The cross-model replication further suggests that while English-centric bias is a shared limitation, intralinguistic sensitivity varies by model, complicating one-size-fits-all solutions.

6. Conclusion

We presented the first systematic evaluation of input language and Portuguese intralinguistic variation on NL2SQL performance. Using translated versions of WikiSQL, we compared two LLMs (LLaMA 3.3 70B and Qwen 3.6 35B, both zero-shot) across English, PT-BR, and PT-PT, keeping schemas, SQL targets, and configuration fixed.

Performance drops markedly from English (47.25% EX) to Portuguese (PT-BR: 40.03%; PT-PT: 39.34%), with the PT-BR vs. PT-PT difference reaching statistical significance ($p = 0.019$). English-only successes outnumber Portuguese-only successes nearly 10:1, and 7.0% of examples show discordant outcomes between Portuguese variants. Stratified analyses reveal that semantic complexity and quantification amplify these effects: `COUNT` queries exhibit the most severe degradation in Portuguese. A replication with Qwen 3.6 35B on a stratified sample confirmed the cross-model robustness of the English-Portuguese gap, while intralinguistic patterns varied, suggesting that the latter are sensitive to model-specific factors such as pretraining data composition.

A multi-layered translation validation with distance metrics, contrastive linguistic analysis, and multi-model judgment confirmed that the translations are semantically faithful (cosine similarity 0.966, semantic equivalence 85.0% PT-BR and 81.0% PT-PT) while exhibiting authentic varietal features (cleft constructions 36.8$\times$ more frequent in PT-PT, enclisis exclusive to PT-PT). The parallel PT-BR/PT-PT WikiSQL corpus will be released for future research.

These findings demonstrate that despite advances in multilingual LLMs, English bias remains a significant challenge for NL2SQL, and intralinguistic variation, while sec-

ondary to language change, produces measurable effects that can reach statistical significance and should not be overlooked in multilingual deployments.

7. Limitations

Four aspects of this study constrain the scope of our conclusions. First, evaluation was restricted to WikiSQL, whose relatively simple queries may not represent complex NL2SQL scenarios involving multi-table joins or nested queries. Second, translations were obtained automatically. Although we conducted extensive validation—distance metrics, contrastive linguistic analysis, and multi-model judgment—these methods do not replace human evaluation, and subtle artifacts may account for part of the PT-BR/PT-PT differences. Third, while both LLaMA 3.3 70B and Qwen 3.6 35B exhibited consistent English-Portuguese gaps, intralinguistic patterns differed between them. Fourth, absolute performance remains below the state of the art for fine-tuned NL2SQL models. Across all dimensions, future work should pursue human-validated parallel corpora, evaluations on complex benchmarks such as Spider, and replication across a broader set of model families.

8. Ethical Considerations

This work uses publicly available datasets (WikiSQL) and does not involve human subjects, proprietary data, or deployment in production systems. The translations were generated automatically and validated through automated methods; no human annotators were employed. We plan to release the parallel PT-BR/PT-PT corpus to foster research on Portuguese NLP, with no restrictions beyond attribution. We do not foresee direct harmful applications of this work, though we acknowledge that NL2SQL systems may perpetuate biases present in pretraining data, including asymmetries between language varieties.

9. Generative AI Use Declaration

Large language models were used as experimental tools in this study: for translating the WikiSQL benchmark into PT-BR and PT-PT (LLaMA 3.3 70B), for generating SQL queries from natural language questions (LLaMA 3.3 70B and Qwen 3.6 35B), and for validating translation quality through multi-model judgment (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B). Additionally, generative AI was used in the preparation of approximately 25% of the manuscript, primarily for polishing and condensing text drafted by the authors. The authors reviewed and edited all AI-generated content and remain fully responsible for the final manuscript.

Acknowledgments

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brasil. Process Number 2025/03245-1.

References

Azevedo, M. M. (2005). *Portuguese: A linguistic introduction*. Cambridge University Press.

Chen, M. et al. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Fan, J., Gu, Z., Zhang, S., Zhang, Y., Chen, Z., Cao, L., Li, G., Madden, S., Du, X., and Tang, N. (2024). Combining small language models and large language models for zero-shot nl2sql. *Proceedings of the VLDB Endowment*, 17(11):2750–2763.

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

Liu, X., Shen, S., Li, B., Ma, P., Jiang, R., Zhang, Y., Fan, J., Li, G., Tang, N., and Luo, Y. (2024). A survey of nl2sql with large language models: Where are we, and where are we going? *arXiv preprint arXiv:2408.05109*.

Meta AI (2024). Llama 3.3 70b model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md`. Accessed: 2025-12-10.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual bert? In *Proceedings of ACL*.

Preda, D., Osório, T., and Cardoso, H. L. (2024). Across the atlantic: Distinguishing between european and brazilian portuguese dialects. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 1*, pages 353–363.

Qwen Team (2026). Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Montemagni, S. and Nivre, J., editors, *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Raffel, C. et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Scholak, T., Schucher, N., and Bahdanau, D. (2021). Picard: Parsing incrementally for constrained auto-regressive decoding from language models.

Shi, P., Zhang, R., Bai, H., and Lin, J. (2022). Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*.

Souza, F., Nogueira, R., and Lotufo, R. (2020). BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Touvron, H. et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wagner Filho, J. A., Wilkens, R., Idiart, M., and Villavicencio, A. (2018). The brWaC corpus: A new open resource for Brazilian Portuguese. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Wang, B. et al. (2020). Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. In *Proceedings of ACL*, pages 7567–7578.

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

Yu, T., Zhang, R., Yang, K., et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of EMNLP*, pages 3911–3921.

Zhao, P., Shao, M., Cui, J., and Zhou, H. (2025). A survey of nl2sql research: From foundations to frontiers. In *2025 2nd International Conference on Intelligent Perception and Pattern Recognition (IPPR)*, pages 396–402. IEEE.

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

Appendix: Prompts Used in Experiments

Translation Prompt (LLaMA 3.3 70B)

System:

> You are a professional translator. Your task is to translate English questions into [Brazilian Portuguese (PT-BR) / European Portuguese (PT-PT)] with high fidelity. Preserve the exact meaning. Do NOT add explanations, comments, markdown, or formatting. Do NOT change numbers, years, names, or entities. Translate as literally as possible while keeping natural grammar. Output ONLY the translated text.
>
> **User:**
>
> Translate the following English question into [Brazilian Portuguese (PT-BR) / European Portuguese (PT-PT)]: "[QUESTION]" Return ONLY the translated text.

NL2SQL Prompt (LLaMA 3.3 70B and Qwen 3.6 35B)

System:

> You are an NL2SQL model specialized in the WikiSQL dataset. Convert a natural language question and a table schema into a structured query in WikiSQL JSON format. Output ONLY a JSON object with fields "sel" (column index), "agg" (0=None, 1=MAX, 2=MIN, 3=COUNT, 4=SUM, 5=AVG), and "conds" (list of [col_idx, op_idx, value] where op_idx: 0==, 1=¿, 2=¡, 3=OP). Return ONLY the JSON, no explanation.
>
> **User:**
>
> Natural language question: [QUESTION]
> Table schema (column indices and names): - [0] col_name_0 - [1] col_name_1 ...
> Some example rows: - row 0: col_name_0=value, col_name_1=value, ...
> Produce ONLY the JSON with fields "sel", "agg", and "conds".

Judge Prompt (DeepSeek V4 Flash, Gemma 4 31B, Qwen 3.6 35B)

System:

You are an expert linguist evaluating translations from English to Portuguese. Assess the translation along three criteria. Answer ONLY with a valid JSON object.

User:

Evaluate the following translation:
ORIGINAL ENGLISH: "[QUESTION_EN]"
TRANSLATION ([PT-BR / PT-PT]): "[QUESTION_PT]"
Criteria: 1. SEMANTIC EQUIVALENCE: Does the translation preserve the exact meaning? (YES / PARTIAL / NO) 2. VARIETAL NATURALNESS: Does it sound natural for a native speaker of [PT-BR / PT-PT]? (NATURAL / ACCEPTABLE / UNNATURAL) 3. COMPLEXITY PRESERVATION: Are interrogative structures, negations, and quantifiers preserved? (YES / PARTIAL / NO)
Return ONLY: {"semantic_equivalence": "...", "varietal_naturalness": "...", "complexity_preservation": "..."}