Language Effects in Text-to-SQL Across English and Portuguese

Lucas Nobre[1], Suele Sousa[1], Savio Teles[1], Anderson Soares[1]

[1]Instituto de Informática - Universidade Federal de Goiás

lucasnobre212@discente.ufg.br

sousa.suele@discente.ufg.br

savioteles@ufg.br

andersonsoares@ufg.br

Abstract

Text-to-SQL systems allow users to query relational databases using natural language, but accuracy remains sensitive to the choice of language, model architecture, and prompting strategy. Although recent Large Language Models (LLMs) incorporate reasoning mechanisms that improve multi-step problem solving in other domains, their effects on multilingual Text-to-SQL are not yet well understood. This work evaluates a diverse set of LLMs on the BIRD benchmark and BIRD_PT, a Portuguese version produced by translating the questions and external knowledge while keeping the original English database schema and values unchanged. We compare four controlled scenarios that vary internal reasoning and guided reasoning for SQL generation. The results show a consistent decrease in accuracy when switching from English to Portuguese, with large variations in robustness across models. Reasoning alone does not reliably improve execution accuracy and can reduce performance in Portuguese, while combining reasoning with a guided plan provides the most stable improvements, although still weaker than in English. These findings highlight ongoing challenges in multilingual Text-to-SQL and emphasize the need to jointly consider language understanding, reasoning activation, and task-aligned planning when designing future systems.

1 Introduction

Accessing information from relational databases often requires writing Structured Query Language (SQL) queries, a skill that demands both technical training and familiarity with the database schemas. For non-technical users, these requirements create a barrier to data-driven decision-making. Text-to-SQL (NL2SQL) addresses this problem by enabling users to interact with databases through natural language, lowering the entry barrier and making data exploration more intuitive and inclusive (Shi et al., 2024; Lin et al., 2024).

As Large Language Models (LLMs) have rapidly advanced, their ability to interpret the nuances and semantic intent of user questions and perform schema understanding has positioned them as the foundation of modern Text-to-SQL systems (Katsogiannis-Meimarakis and Koutrika, 2023). Contemporary approaches now rely heavily on prompting, in-context learning, and task-specific reasoning strategies to generate SQL queries that are both structurally valid and semantically aligned with the user's intent (Hong et al., 2025). Recent work further augments LLMs with explicit reasoning mechanisms such as chain-of-thought and specialized reasoning modes (LRMs), which have demonstrated strong gains in tasks involving multi-step mathematical, logical, and code reasoning (OpenAI et al., 2024; DeepSeek-AI et al., 2025; Team, 2025; Princis et al., 2025). Since Text-to-SQL also requires decomposing user intent, linking schema elements, and composing executable queries, it is natural to expect that these reasoning capabilities could translate into improved Text-to-SQL performance.

However, most of the progress in Text-to-SQL has been measured on English-centric benchmarks and in monolingual settings. Multilingual LLMs are still constrained by English-heavy pretraining corpora, which leads to weaker coverage and noisier representations in other languages (Cruz-Castañeda and Amadeus, 2025; Pham et al., 2025). In multilingual scenarios, models often retain strong internal reasoning abilities but fail to reliably activate them when comprehension in the target language is incomplete, causing errors rooted in language understanding rather than reasoning itself (Huang et al., 2024; Shi et al., 2022). These limitations are amplified in Text-to-SQL, where precise alignment between natural language, database schemas, and SQL operators is critical. For Portuguese, prior work has shown that reduced lexical overlap with SQL structures, richer morphology,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 270–280

and domain-specific terminology make schema linking and clause interpretation particularly challenging, resulting in lower accuracy than in English (José et al., 2022; Pedroso et al., 2025; José and Cozman, 2021).

It is still unclear how modern reasoning mechanisms interact with these multilingual bottlenecks. On the one hand, explicit reasoning and guided plans could provide a stable scaffold that helps models overcome noisy language understanding and produce more reliable SQL. On the other hand, reasoning traces and system-level thinking modes are typically optimized and evaluated in English, which may introduce additional friction when prompts, questions, and external knowledge are presented in another language. As a result, it is not obvious whether activating reasoning will narrow or widen the performance gap between English and Portuguese Text-to-SQL, nor which model families are more robust to this shift.

In this work, we provide a systematic empirical study of Text-to-SQL in Portuguese under modern LLM-based settings. We build on the BIRD benchmark (Li et al., 2023), a large-scale, cross-domain Text-to-SQL dataset, and construct BIRD_PT, a Brazilian Portuguese version obtained by translating the questions and external knowledge with a state-of-the-art LLM. Using BIRD and BIRD_PT, we evaluate a diverse set of proprietary and open-weight models under four controlled scenarios that independently toggle internal reasoning and a guided SQL generation plan. The plan is a short natural-language template that outlines the steps for constructing the SQL query. This design allows us to disentangle the effects of language, reasoning, planning, and model architecture on execution accuracy.

Overall, the contributions of this work are: (i) a large-scale evaluation of reasoning and guided planning for Text-to-SQL in both English and Portuguese, (ii) an empirical characterization of how different LLM families trade off accuracy and cross-lingual robustness in Text-to-SQL, and (iii) BIRD_PT, a new Portuguese counterpart of a widely used Text-to-SQL benchmark. These findings highlight that multilingual Text-to-SQL remains far from solved and that simply "turning on" reasoning is insufficient: effective systems must jointly account for language, reasoning mechanisms, and task-aligned guidance.

2 Related Work

Text-to-SQL systems. Early Text-to-SQL research relied on rule-based grammars and semantic parsing, but these approaches required extensive manual engineering and did not generalize well across domains (Affolter et al., 2019; Guo et al., 2019). With deep learning, sequence-to-sequence models improved semantic parsing (Zhong et al., 2017), and encoder–decoder architectures introduced relation-aware attention to better align questions and schema elements (Wang et al., 2020). Recent progress is dominated by LLM-based approaches that combine semantic understanding with schema grounding through prompting and in-context learning in zero-shot (Dong et al., 2023) and few-shot settings (Brown et al., 2020). Modern systems use schema linking, task decomposition, and self-correction within modular pipelines (Pourreza and Rafiei, 2023; Pourreza et al., 2025; Shkapenyuk et al., 2025; Li et al., 2024; Fu et al., 2023), further extended by multi-agent coordination (Wang et al., 2025a), ensemble-based generation (Liu et al., 2025b), and execution-guided refinement (Talaei et al., 2024). Other studies rely on supervised fine-tuning (SFT) (Lee et al., 2025; Li et al., 2025), which fine-tunes open-source LLMs using public Text-to-SQL datasets.

Reasoning Models. Recent progress shows that reinforcement learning and chain-of-thought prompting enhance multi-step reasoning in domains such as mathematics, logic, and code generation, exemplified by (OpenAI et al., 2024; DeepSeek-AI et al., 2025; Team, 2025; Princis et al., 2025). Since Text-to-SQL similarly requires decomposing user intent, linking schema elements, and generating executable queries, reasoning-based methods have been adopted to improve its compositional accuracy (Yao et al., 2025; Papicchio et al., 2025; Liu et al., 2025a; Ali et al., 2025). Recent evidence shows that reasoning alone is not sufficient to improve performance in the Text-to-SQL task and that reasoning must be guided by task-aligned structure to be effective (Papicchio et al., 2025).

Multilingual LLMs, reasoning, and Text-to-SQL in Portuguese. Multilingual LLMs face limitations because English-centric pretraining leads to weaker coverage in other languages, which becomes especially problematic in tasks requiring precise alignment between natural language and structured information (Cruz-Castañeda and Amadeus, 2025; Pham et al., 2025). In multilingual reason-

ing, models often retain strong internal reasoning capabilities but fail to reliably activate them when comprehension in the target language is incomplete, leading to errors rooted in language understanding rather than reasoning itself (Huang et al., 2024; Shi et al., 2022). Recent work shows that multilingual LLMs often default to English during internal reasoning, regardless of the input language (Wang et al., 2025b). These characteristics directly affect Text-to-SQL in Portuguese, where reduced lexical overlap with SQL structures and greater morphological variability make schema linking and clause interpretation more difficult, resulting in less stable SQL generation than in English (José et al., 2022; Pedroso et al., 2025; José and Cozman, 2021). Our work explicitly examines how this multilingual bottleneck interacts with reasoning and guided reasoning by contrasting English BIRD with its Portuguese BIRD_PT counterpart.

3 Methodology

Our approach aims to investigate how using Portuguese impacts the results in the Text-to-SQL task and how reasoning mechanisms affect performance. We structure our methodology around the (i) impact of translating user questions and the prompt template from English to Portuguese, (ii) the contribution of reasoning and guided plans to SQL generation, and (iii) the robustness of different model families under multilingual conditions. To accomplish this, we treat reasoning, planning, and language as independent experimental factors applied systematically across different model families.

The Text-to-SQL task is defined as a function $f : (q, S, K) \rightarrow$ SQL that maps a natural-language question q, a database schema S, and optional external knowledge K to a SQL query whose execution $\text{exec}(\text{SQL}, D)$ matches the gold answer. We evaluate our results using the `Execution Accuracy` metric. Execution accuracy evaluates whether the SQL query predicted by the model, when executed on the database, produces the same results as the reference SQL. We also make use of a *plan*, which is a short natural-language description of the steps involved in forming the SQL query. The plan is included in the prompt as lightweight guidance during generation.

To evaluate reasoning and planning, we define four different scenarios. Each scenario corresponds to a different combination of reasoning or no reasoning, and the presence or absence of a guided plan, allowing us to evaluate both factors independently and jointly. To understand the effect of using Portuguese in the Text-to-SQL task, we translate the BIRD benchmark into Portuguese using `gemini-2.5-pro` along with our prompt templates and plans. We translated only the benchmark questions and domain knowledge; the schema and database values remained in English. To assess the quality of the translated benchmark, we manually reviewed 100 random samples of question-knowledge pairs. We identified only one clear translation error, suggesting that the automatic translation was reliable for our experiments

Across the BIRD benchmark, we evaluated 22 distinct SQL generation plans. All plans were evaluated in all scenarios using Gemini-2.5-Flash. We generated three plans with Claude-4-Sonnet, three with Gemini-2.5-Flash, and four with GPT-4.1. We identified that plans containing query examples achieved higher accuracy, particularly those generated with Gemini-2.5-Flash. After observing these initial results, we generated 12 additional plans using Gemini-2.5-Flash. We then selected the plan with the highest average accuracy among all plans.

We evaluate a diverse set of LLMs spanning proprietary and open-weight families, including models without reasoning, models with optional reasoning modes, and models whose reasoning is always enabled. The proprietary group includes `gemini-2.5-flash`, `gemini-2.5-pro`, `claude-4.5-Sonnet`, and `claude-4.5-Haiku`, while the open-weight group covers `deepseek-r1`, `deepseek-v3.1`, `qwen3-coder`, `qwen3-next`, and `qwen3-next-thinking`. This selection enables controlled comparisons across architectures, reasoning behaviors, and multilingual robustness.

We ran our experiments with all combinations of models, benchmarks, and scenarios, allowing us to isolate the effects of language, reasoning, planning, and model architecture. All the code used in our experiments is available in `https://anonymous.4open.science/r/text_to_sql_eval-8855/README.md`

3.1 Qualitative Analysis Procedure

The objective of the qualitative analysis is to understand the main failure modes that underlie the English-Portuguese performance gap in Text-to-SQL, distinguishing errors caused by translation artifacts from those arising from language comprehension, schema grounding, and SQL generation.

In addition to aggregate accuracy metrics, we

```
1. Deconstruction of the Request:
   - Target Entity: <Target_Table>
   - Output Attributes: <Column_1>, <
       Column_2>
   - Filtering Criterion: <Filter_Column
       > = 'Filter_Value'
   - Sorting Criterion: <Sort_Column>

2. SQL Structure:
   SELECT <Column_1>, <Column_2>
   FROM <Target_Table>
   WHERE <Filter_Column> = 'Filter_Value
       '
   ORDER BY <Sort_Column> ASC/DESC

3. Generation Instruction:
   Produce an SQL query following
       exactly the structure above.
```

Figure 1: Deconstruction of the Request – Filtering and Sorting Plan

performed a targeted qualitative error analysis to better understand the sources of performance differences between English and Portuguese. We initially inspected 407 instances of all scenarios on model `gemini-2.5-flash`. For each case, we examined the English and Portuguese outputs side-by-side and manually categorized errors related to translation differences, comprehension shifts, additional or missing columns, and filtering inconsistencies.

4 Experimental Setup

4.1 Plans

We first conducted a systematic plan search on the BIRD development set, evaluating 22 distinct SQL generation plans with `gemini-2.5-flash` across all four scenarios and both languages. For each candidate, we computed the average execution accuracy over English and Portuguese, with and without reasoning. The "Deconstruction of the Request – Filtering and Sorting" plan emerged as the best overall option, achieving the highest average accuracy and consistently strong performance in non-reasoning settings while remaining competitive in reasoning scenarios.

For our experimental setup, we therefore employ this plan as the default guided template for all models. The plan is written in English when using the BIRD benchmark and in Portuguese when using the BIRD_PT benchmark. The plan is shown in Figure 1.

4.2 Scenarios

We define four scenarios to evaluate LLM performance on the Text-to-SQL task, using the notation Reasoning (R) and Planning (P): (i) No reasoning, no planning(**–/–**), (ii) No reasoning, with planning (**–/P**), (iii) reasoning, no planning (**R/–**), and (iv) reasoning with planning (**R/P**). In all cases, the prompt instructs the model to generate a SQL query that answers the user's question using the provided database schema and any available external knowledge. For the scenarios that include reasoning, the model's internal reasoning mechanism (when supported) is enabled with a maximum budget of 1,024 reasoning tokens. For the scenarios without reasoning, this internal mechanism is disabled. The **no reasoning, no planning** scenario performs direct SQL generation with neither internal reasoning nor a predefined plan. The **no reasoning, with planning** scenario guides the SQL generation using a predefined plan, but internal reasoning remains disabled. The **reasoning, no planning** uses internal reasoning but without a predefined plan, while **reasoning with planning** combines a predefined plan with internal reasoning when generating the SQL.

Not all model–scenario combinations are supported by the underlying APIs. Models that do not have a thinking mode (*e.g.*, `qwen3-coder-480b-a3b-instruct`, `qwen3-next-480b-a3b-instruct`) are only evaluated in the non-reasoning scenarios (**–/–** and **–/P**). Conversely, models whose thinking mode is always enabled (*e.g.*, `gemini-2.5-pro`, `deepseek-r1-0528`, `qwen3-next-480b-a3b-thinking`) are only evaluated in the reasoning scenarios (**R/–** and **R/P**). In all tables, unsupported model–scenario pairs are marked with "–".

4.3 Benchmarks

We evaluate our scenarios using the challenging Text-to-SQL benchmark BIRD (Li et al., 2023) and a translated version into Brazilian Portuguese that we call BIRD_PT. BIRD is a large-scale, cross-domain Text-to-SQL dataset that contains 95 databases spanning 37 professional domains. Designed to closely relate to real-world databases, the data contains large, dirty, and noisy values, making value comprehension a challenge. In this paper, we use the development set for evaluation, which contains 1,534 Text-to-SQL pairs from 11 databases. Each pair has a question, the ground-truth SQL, and the optional external knowledge containing domain or database information necessary to create an accurate SQL query. BIRD_PT is a translated version of the BIRD benchmark we developed for this study. It has all the development

set instances translated into Brazilian Portuguese using `gemini-2.5-pro`. Only the questions and external knowledge were translated into Portuguese. When using the Portuguese benchmark, the generated SQL query, plans, knowledge, and questions keep the standard SQL keywords in English (*e.g.*, SELECT, FROM, WHERE), as SQL syntax is not subject to translation. The BIRD_PT benchmark, along with the evaluation code and experimental scripts, is available in the linked GitHub repository

4.4 Models

We use a variety of contemporary models to understand the impacts of different sizes, architectures, training methods, and reasoning capabilities. Our selection includes both proprietary and open-weight models, as well as models with thinking mode available, models that do not have thinking mode, and models that only have thinking mode.

The proprietary models include `gemini-2.5-flash`, `gemini-2.5-pro` (Comanici et al., 2025), `claude-4.5-Sonnet` (Anthropic, 2025b), and `claude-4.5-Haiku` (Anthropic, 2025a). `gemini-2.5-flash` and both Claude models have thinking modes that we can enable or disable, while `gemini-2.5-pro` always has its thinking enabled.

The open-source models we used consist of the Deepseek and Qwen families: `deepseek-r1-0528` (DeepSeek-AI et al., 2025), `deepseek-v3.1` (DeepSeek-AI et al., 2025), `qwen3-coder-480b-a3b-instruct`, `qwen3-next-480b-a3b-instruct` and `qwen3-next-480b-a3b-thinking` (Team, 2025). The Deepseek-R1 and Qwen3-Next-Thinking models are thinking models, so their thinking mode is always enabled.

4.5 Implementation Details

We use the LiteLLM package (BerriAI, 2024) with Vertex AI to run the LLM models that we selected. Across all models, we use `temperature=0`, except for Claude models, which require `temperature=1` to enable thinking mode, according to LiteLLM API specifications. All scenarios use the same prompt template for generating SQL, which includes the database schema associated with the question, the external knowledge, if available, the selected plan, and the instruction to generate a SQL query based on the information provided.

5 Results

In this section, we present a comprehensive evaluation of our experimental results. We organize the analysis into four parts: (i) the impact of switching from English (BIRD) to Portuguese (BIRD_PT), (ii) the effect of guided plans, (iii) the effect of internal reasoning, and (iv) differences across model families and sizes. Table 1 reports detailed results for all experimental configurations.

5.1 Text-to-SQL in Portuguese and English

To quantify the effect of language on Text-to-SQL accuracy, we compare model performance on the BIRD benchmark and its Portuguese counterpart, BIRD_PT, using the same experimental configurations. Across all models and scenarios, accuracy consistently decreases when moving from English to Portuguese, confirming that language choice has a measurable impact on Text-to-SQL performance.

The magnitude of this drop varies considerably by model. Proprietary models such as `gemini-2.5-pro` and `claude-sonnet-4.5` exhibit the smallest drops, at 0.8pp and 3.0pp respectively. Among open-weight models, `qwen3-next-480b` is notably strong, with a relatively small gap of 2.6pp, outperforming other open-weight systems such as `deepseek-v3.1` (4.2pp) and `qwen3-coder-480b` (4.7pp). In contrast, `deepseek-r1` suffers from particularly severe degradation under Portuguese reasoning with plan conditions. The experiments revealed difficulties in understanding the questions and frequent failures to follow the instructions in the prompt, sometimes adding irrelevant columns or relying solely on external knowledge to answer the question.

We observed in our experiments that all models, except `deepseek-r1`, produced their thinking tokens exclusively in English, even when the question, plan, and knowledge were entirely in Portuguese. This behavior, known as language mixing, reveals a limitation in the multilingual reasoning capabilities of current LLMs, where they can understand the Portuguese inputs well but still use English in their internal reasoning. It suggests that the models utilize part of their capability to align the Portuguese question with English thinking, degrading the overall task performance.

A qualitative inspection of 407 examples shows that many Portuguese errors arose from misinterpreting question details, filtering conditions with altered casing or additional terms, and from generating unnecessarily complex SQL queries. These issues are much less frequent in English and help explain the systematic performance drop in Portuguese. Even when the English and Portuguese

Table 1: Comprehensive performance comparison across BIRD (EN) and BIRD_PT (PT) benchmarks. Values show accuracy (%). Δ represents performance gap (EN - PT) in pp.

	BIRD (EN)					BIRD_PT (PT)					
Model	**–/–**	**–/P**	**R/–**	**R/P**	**Avg**	**–/–**	**–/P**	**R/–**	**R/P**	**Avg**	**Δ**
claude-haiku-4.5	51.1	50.8	52.5	54.5	52.2	46.9	46.7	47.2	51.0	48.0	-4.3
claude-sonnet-4.5	63.9	63.6	60.9	**65.0**	63.4	**61.9**	61.8	57.3	60.7	60.4	-3.0
gemini-2.5-flash	62.6	63.6	60.5	62.3	62.2	59.2	60.4	57.1	57.8	58.6	-3.6
gemini-2.5-pro	–	–	62.1	62.5	62.3	–	–	60.5	**61.9**	61.5	-0.8
deepseek-r1	–	–	53.8	48.5	50.0	–	–	48.1	21.3	25.1	-24.9
deepseek-v3.1	57.8	59.0	54.3	59.9	57.8	55.3	55.2	48.6	54.9	53.5	-4.2
qwen3-coder-480b	60.4	61.3	–	–	61.2	57.0	56.8	–	–	56.5	-4.7
qwen3-next-480b	61.2	61.5	–	–	61.4	59.5	57.5	–	–	58.7	-2.6
qwen3-next-480b-thinking	–	–	57.9	58.0	58.0	–	–	52.5	56.3	54.3	-3.6

R: Reasoning, P: Planning. –/–: No reasoning, no planning. –/P: No reasoning, with planning. R/–: Reasoning, no planning. R/P: Reasoning with planning. "–" indicates model–scenario combinations that the underlying API does not support (no reasoning mode or reasoning always on), so these settings are not evaluated.

Table 2: Effect of translation difference found

Model: gemini-2.5-flash

Question	**Gold SQL Snippet**	**Generated SQL Snippet**
What is the unabbreviated mailing street address of the school with the highest FRPM count for K-12 students?	`SELECT T2.MailStreet FROM frpm`	`SELECT T1.MailStreet FROM schools`
Qual é o endereço postal completo da escola com o maior FRPM count para alunos K-12?	`SELECT T2.MailStreet FROM frpm`	`SELECT T2.MailStreet, T2.MailCity, T2.MailZip, T2.MailState FROM frpm`

Question – Item from BIRD. Gold SQL Snippet – Correct SQL fragment. Generated SQL Snippet – Fragment where the predicted SQL differs from Gold SQL.

prompts were semantically equivalent, models handled the Portuguese inputs less consistently during SQL generation. In all cases analyzed, no significant quantities were found where the translation was considered incorrect or dubious. Only one observed case suggested that the translation may have induced the model's error, as can be seen in Table 2.

The most common errors observed are related to filter applications, which may explain the differences between BIRD_PT and BIRD. Specifically, some models employed uppercase characters even when the provided external knowledge suggested otherwise. In addition, misinterpretation of the question frequently led to the inclusion of unnecessary fields or additional filters, as can be seen in Table 3.

The results show us two key findings: (i) the performance drop when shifting to Portuguese is systematic, affecting all evaluated models. (ii) The gap magnitude is model-dependent, suggesting that some models transfer question understanding and schema reasoning capabilities more reliably across languages.

5.2 Effect of Guided Plan

Across English and Portuguese, the inclusion of the guided plan consistently improved performance in reasoning scenarios and had a negligible effect on scenarios without reasoning for most models, as shown in Table 4. Given these results, we restrict our analysis to the reasoning scenarios. In these scenarios, using the plan led to substantial accuracy gains, particularly for `deepseek-v3.1`, improving by 5.6pp on BIRD and 6.3pp on BIRD_PT, and for `claude-sonnet-4.5`, which improves by 4.1pp on BIRD and 3.4pp on BIRD_PT. In contrast to these gains, the guided plan decreased `deepseek-r1` performance by 5.3pp on BIRD and by 26.8pp on BIRD_PT, representing the largest negative delta among all evaluated models. This outcome is consistent with the limitations described in the Deepseek-R1 system report, which states that the model is optimized for Chinese and English (DeepSeek-AI et al., 2025).

In our experiments, we used a universal plan for all questions, even though in real-world scenarios, we use different strategies to solve different problems. A problem may benefit from a join-focused decomposition strategy, while other problems might benefit from a simpler relational-filtering plan. An interesting direction for future work is to design adaptive plans that dynamically provide a strategy most appropriate for the particular SQL task at hand.

Table 3: Differences found in Portuguese questions and predicted SQL.

Model: gemini-2.5-flash

Question	Gold SQL Snippet	Generated SQL Snippet
List all the card id and artist with unknown power which are legal for commander play format.	`WHERE T2.format = 'commander'`	`WHERE T2.format = 'commander'`
Liste o card id e o artista das cartas com power desconhecido que são legais para o formato Commander.	`T2.format = 'commander'`	`T2.format = 'Commander'`
Please list the lowest three eligible free rates for students aged 5–17 in continuation schools.	`SELECT 'Free Meal Count (Ages 5-17)' / 'Enrollment (Ages 5-17)' FROM frpm`	`SELECT CAST(T1.'Free Meal Count (Ages 5-17)' AS REAL) * 1.0 / T1.'Enrollment (Ages 5-17)' FROM frpm AS T1`
Por favor, liste as três menores taxas de gratuidade elegíveis para alunos de 5 a 17 anos em escolas de ensino supletivo.	`SELECT 'Free Meal Count (Ages 5-17)' / 'Enrollment (Ages 5-17)' FROM frpm`	`SELECT T1.'School Name', CAST(T1.'Free Meal Count (Ages 5-17)' AS REAL) / T1.'Enrollment (Ages 5-17)' FROM frpm`

Question – Item from BIRD. Gold SQL Snippet – Correct SQL fragment. Generated SQL Snippet – Fragment where the predicted SQL differs from Gold SQL.

Overall, the results show that introducing a guided plan in reasoning scenarios has a positive and language-agnostic effect. The plan elicits task-specific reasoning for Text-to-SQL, guiding the model reasoning process during SQL generation and thus improving accuracy. This indicates that many models do not naturally activate reasoning patterns that are well aligned with SQL problem-solving.

In contrast, the guided plan produced no meaningful performance change in non-reasoning scenarios. In our experiments, the model is instructed to output only the final SQL query, without generating intermediate steps or explanations. Because the model is not permitted to externalize any step-by-step reasoning (Wei et al., 2022), it cannot integrate the plan's steps into its reasoning process. This highlights that planning is most useful when the model can externalize intermediate steps.

Table 4: Effect of planning under no-reasoning (NR) and reasoning (R) for BIRD (EN) and BIRD_PT (PT). Values show accuracy (%). Δ represents performance gap (NoP - P) in pp.

	NR		R		Δ	
Model	**NoP**	**P**	**NoP**	**P**	**NR**	**R**
BIRD (EN)						
c-haiku-4.5	51.1	50.8	52.5	54.5	-0.3	+2.0
c-sonnet-4.5	63.9	63.6	60.9	65.0	-0.3	+4.1
g-2.5-flash	62.6	63.6	60.5	62.3	+1.0	+1.8
g-2.5-pro	–	–	62.1	62.5	–	+0.4
deepseek-r1	–	–	53.8	48.5	–	-5.3
deepseek-v3.1	57.8	59.0	54.3	59.9	+1.2	+5.6
qwen3-coder	60.4	61.3	–	–	+0.9	–
qwen3-next	61.2	61.5	–	–	+0.3	–
qwen3-next-t	–	–	57.9	58.0	–	+0.1
Avg	59.50	59.97	57.43	58.67	+0.47	+1.24
BIRD_PT (PT)						
c-haiku-4.5	46.9	46.7	47.2	51.0	-0.2	+3.8
c-sonnet-4.5	61.9	61.8	57.3	60.7	-0.1	+3.4
g-2.5-flash	59.2	60.4	57.1	57.8	+1.2	+0.7
g-2.5-pro	–	–	60.5	61.9	–	+1.4
deepseek-r1	–	–	48.1	21.3	–	-26.8
deepseek-v3.1	55.3	55.2	48.6	54.9	-0.1	+6.3
qwen3-coder	57.0	56.8	–	–	-0.2	–
qwen3-next	59.5	57.5	–	–	-2.0	–
qwen3-next-t	–	–	52.5	56.3	–	+3.8
Avg	56.63	56.40	53.04	52.0	-0.23	-1.05

NR = No Reasoning. R = Reasoning. NoP = No Plan. P = With Plan. "–" indicates unavailable settings.
"–" indicates model–scenario combinations that the underlying API does not support (no reasoning mode or reasoning always on), so these settings are not evaluated.

5.3 Effect of Reasoning

To understand the effect of reasoning, we analyze the results between reasoning scenarios and non-reasoning scenarios. Across BIRD and BIRD_PT, reasoning yielded heterogeneous results on the models. The results are in Table 5. The `claude-haiku-4.5` (`c-haiku-4.5`) model yields the most positive effect, increasing the accuracy by 2.5pp on BIRD and 2.5pp on BIRD_PT, followed by `qwen3-next-480b` (`qwen3-next`) with 0.7pp and 0.6pp respectively. The `deepseek-v3.1` had the worst drop in accuracy, with -1.4pp on BIRD and -3.3pp on BIRD_PT, followed by `gemini-2.5-flash` (`g-2.5-flash`) with -1.7pp on BIRD and -3.0pp on BIRD_PT. On average, reasoning scenarios had worse results than non-reasoning scenarios for both benchmarks.

These results indicate that activating reasoning does not universally improve SQL generation, however, as discussed in 5.2, enabling reasoning and

Table 5: Effect of reasoning on total execution accuracy across models for BIRD (EN) and BIRD_PT (PT). Values shown in accuracy (%). Δ represents performance gap (NoR - R) in pp.

	BIRD (EN)			BIRD_PT (PT)		
Model	**NoR**	**R**	**Δ**	**NoR**	**R**	**Δ**
c-haiku-4.5	51.1	53.6	+2.5	46.7	49.2	+2.5
c-sonnet-4.5	63.7	62.8	-0.8	61.2	59.5	-1.7
g-2.5-flash	63.2	61.5	-1.7	60.0	57.1	-2.7
g-2.5-pro	–	62.6	–	–	61.0	–
deepseek-r1	–	48.9	–	–	32.2	–
deepseek-v3.1	58.1	56.7	-1.4	55.1	51.8	-3.3
qwen3-coder	60.8	–	–	57.1	–	–
qwen3-next	62.0	62.6	+0.7	59.0	59.6	+0.6
qwen3-next-t	–	58.1	–	–	53.7	–
Average	59.1	58.2	-0.9	56.0	53.0	-3.0

NoR: Without reasoning. R: With reasoning. "–" indicates the model only provides one of the two settings.
"–" indicates model–scenario combinations that the underlying API does not support (no reasoning mode or reasoning always on), so these settings are not evaluated.

using plans shows improvements over scenarios without reasoning. We can see that the language used does not shift the trend, but it lowers the accuracy gains for all models.

Taken together, these findings reveal two insights: (i) Internal reasoning abilities alone are insufficient for reliable performance gains in Text-to-SQL and must be complemented by task-aligned structures, such as guided plans. (ii) The results in Portuguese when using reasoning vary by model, with some models showing no impact, while others experience a significant drop in performance.

5.4 Effect of Model Family and Sizes

The results reveal clear differences across model families in their ability to maintain performance when shifting from English to Portuguese. Claude and Gemini models achieve the strongest and most stable results overall, with high average accuracies and the smallest cross-lingual gaps.

Qwen models perform competitively, though there are moderately larger gaps between English and Portuguese. Both Qwen3-Next models and Deepseek-v3.1 rank highest among the open-source options, but their average drops of 2.6pp and 4.7pp suggest that multilingual capabilities are less stable than those of Claude and Gemini.

Model size contributes to performance but does not solely determine it. Gemini-2.5-flash and Gemini-2.5-pro have similar results, despite the flash variant being smaller than the pro. Overall, these findings show that architectural design and multilingual training are the primary drivers of cross-lingual robustness in Text-to-SQL systems.

6 Limitations

We acknowledge that our study has limitations. The Portuguese benchmark was generated through the automatic translation of the questions and external knowledge in the BIRD benchmark using the Gemini-2.5 Pro model without a full human validation pass, which may introduce subtle linguistic artifacts. The database schema, table names, and their values were kept in English and were not localized to the Brazilian context, potentially underestimating the complexity of multilingual Text-to-SQL settings. A single plan was used for all the experiments, even though different queries may benefit from different instructions and contexts. Addressing these limitations will help better understand the effects of the Portuguese language in Text to SQL tasks.

7 Conclusion

Our study provides a systematic evaluation of modern LLMs performing Text-to-SQL in Portuguese, revealing that shifting from English consistently reduces execution accuracy across all models, with varying magnitudes. We show that internal reasoning alone does not guarantee performance improvements and can even degrade accuracy, particularly in Portuguese, while guided SQL generation plans offer more stable gains, albeit with weaker effects compared to English. Even with Portuguese inputs, reasoning LLMs generate English thinking tokens, introducing an additional source of misalignment in SQL generation. These results demonstrate that multilingual Text-to-SQL remains challenging, as even highly capable LLMs struggle to transfer schema reasoning and SQL composition across languages without task-aligned guidance. Our findings highlight the need for future systems to jointly consider language understanding, reasoning activation, and structured planning.

We identify several directions for future research. First, BIRD_PT was translated automatically, and a human validation pass is needed to establish a high-quality Portuguese benchmark. Second, we kept the database schema and values in English; future work should examine fully localized databases or mixed-language schemas to better understand cross-lingual effects. Third, reasoning models generated their thinking tokens exclusively in English,

even with Portuguese inputs, raising the question of whether inducing Portuguese reasoning could improve schema linking and SQL generation. Finally, because Portuguese is morphologically richer and more syntactically variable than English, developing finer-grained error taxonomies may help isolate the linguistic sources of model failures and guide improvements in multilingual Text-to-SQL.

8 Acknowledgments

This work has been funded by P&D CEMIG/ANEEL PD-04950-D0677/2023. We acknowledge the support provided by the Center of Excellence in Artificial Intelligence (CEIA), affiliated with the Federal University of Goiás (UFG). This work was supported by the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Katrin Affolter, Kurt Stockinger, and Abraham Bernstein. 2019. A comparative survey of recent natural language interfaces for databases. *The VLDB Journal*, 28(5):793–819.

Alnur Ali, Ashutosh Baheti, Jonathan Chang, Ta-Chung Chi, Brandon Cui, Andrew Drozdov, Jonathan Frankle, Abhay Gupta, Pallavi Koppol, Sean Kulinski, Jonathan Li, Dipendra Misra, Krista Opsahl-Ong, Jose Javier Gonzalez Ortiz, Matei Zaharia, and Yue Zhang. 2025. A State-of-the-Art SQL Reasoning Model using RLVR. *Preprint*, arXiv:2509.21459.

Anthropic. 2025a. Claude haiku 4.5 [large language model – system card]. `https://www.anthropic.com/claude-haiku-4-5-system-card`. Accessed: 2025-12-01.

Anthropic. 2025b. Claude sonnet 4.5: System card. `https://www.anthropic.com/claude-sonnet-4-5-system-card`. Accessed: 2025-11-08.

BerriAI. 2024. Litellm: Unified python sdk and proxy server for 100+ llms. Accessed: 2025-12-05.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint*. ArXiv:2507.06261 [cs].

William Alberto Cruz-Castañeda and Marcellus Amadeus. 2025. Large Languages Models in Brazilian Portuguese: A Chronological Survey. *Journal of the Brazilian Computer Society*, 31(1):1168–1187.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Xuemei Dong, Chao Zhang, Yuhang Ge, Yuren Mao, Yunjun Gao, lu Chen, Jinshu Lin, and Dongfang Lou. 2023. C3: Zero-shot Text-to-SQL with ChatGPT. *Preprint*, arXiv:2307.07306.

Han Fu, Feifei Li, Chang Liu, Jian Tan, Bin Wu, and Jianling Sun. 2023. CatSQL: Towards real world natural language to SQL applications. *Proceedings of the VLDB Endowment*, 16(6):1534–1547.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-SQL in cross-domain database with intermediate representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4524–4535, Florence, Italy. Association for Computational Linguistics.

Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. 2025. Next-generation database interfaces: A survey of llm-based text-to-sql. *IEEE Transactions on Knowledge and Data Engineering*, 37(12):7328–7345.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. MindMerger: Efficiently Boosting LLM Reasoning in non-English Languages. *Advances in Neural Information Processing Systems*, 37:34161–34187.

Marcelo Archanjo José and Fabio Gagliardi Cozman. 2021. mrat-sql+gap: A portuguese text-to-sql transformer. In *Intelligent Systems*, pages 511–525, Cham. Springer International Publishing.

Marcos Menon José, Marcelo Archanjo José, Denis Deratani Mauá, and Fábio Gagliardi Cozman. 2022. Integrating question answering and text-to-sql in portuguese. In *Computational Processing of the Portuguese Language*, pages 278–287, Cham. Springer International Publishing.

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

Dongjun Lee, Choongwon Park, Jaehyuk Kim, and Heesoo Park. 2025. MCS-SQL: Leveraging multiple prompts and multiple-choice selection for text-to-SQL generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 337–353, Abu Dhabi, UAE. Association for Computational Linguistics.

Haoyang Li, Shang Wu, Xiaokang Zhang, Xinmei Huang, Jing Zhang, Fuxin Jiang, Shuai Wang, Tieying Zhang, Jianjun Chen, Rui Shi, Hong Chen, and Cuiping Li. 2025. Omnisql: Synthesizing high-quality text-to-sql data at scale. *Proc. VLDB Endow.*, 18(11):4695–4709.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, and 1 others. 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.

Zhishuai Li, Xiang Wang, Jingjing Zhao, Sun Yang, Guoqing Du, Xiaoru Hu, Bin Zhang, Yuxiao Ye, Ziyue Li, Rui Zhao, and Hangyu Mao. 2024. PET-SQL: A Prompt-Enhanced Two-Round Refinement of Text-to-SQL with Cross-consistency. *Preprint*, arXiv:2403.09732.

Jie Lin, Yulong Liang, Jiyan Li, Yi Bai, and Yong Wang. 2024. Nl2sql with partial missing metadata based on multi-view metadata graph compensation and reasoning. *Applied Intelligence*, 54(2):1511–1524.

Shu Liu, Alan Zhu, Sumanth Hegde, Shiyi Cao, Shuo Yuan, Samion Suwito, Tyler Griggs, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025a. SkyRL-SQL: Multi-turn SQL Data Agents via RL. In *First Workshop on Multi-Turn Interactions in Large Language Models*.

Yifu Liu, Yin Zhu, Yingqi Gao, Zhiling Luo, Xiaoxia Li, Xiaorong Shi, Yuntao Hong, Jinyang Gao, Yu Li, Bolin Ding, and Jingren Zhou. 2025b. Xiyan-sql: A novel multi-generator framework for text-to-sql. *Preprint*, arXiv:2507.04701.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carneya, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Simone Papicchio, Simone Rossi, Luca Cagliero, and Paolo Papotti. 2025. Think2sql: Reinforce llm reasoning capabilities for text2sql. *Preprint*, arXiv:2504.15077.

Breno Carvalho Pedroso, Marluce Rodrigues Pereira, and Denilson Alves Pereira. 2025. Performance evaluation of LLMs in the Text-to-SQL task in Portuguese. In *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, pages 260–269. SBC.

Khanh Trinh Pham, Thu Huong Nguyen, Jun Jo, Quoc Viet Hung Nguyen, and Thanh Tam Nguyen. 2025. Multilingual Text-to-SQL: Benchmarking the Limits of Language Models with Collaborative Language Agents. *arXiv preprint*. ArXiv:2509.24405 [cs].

Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan Arik. 2025. Chase-sql: Multi-path reasoning and preference optimized candidate selection in text-to-sql. In *International Conference on Representation Learning*, volume 2025, pages 60385–60415.

Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: decomposed in-context learning of text-to-sql with self-correction. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Henrijs Princis, Cristina David, and Alan Mycroft. 2025. Enhancing SQL Query Generation with Neurosymbolic Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19959–19968.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. https://arxiv.org/abs/2210.03057v1.

Liang Shi, Zhengju Tang, Nan Zhang, Xiaotong Zhang, and Zhi Yang. 2024. A survey on employing large language models for text-to-sql tasks. *ACM Computing Surveys*.

Vladislav Shkapenyuk, Divesh Srivastava, Theodore Johnson, and Parisa Ghane. 2025. Automatic Metadata Extraction for Text-to-SQL. *arXiv preprint*. ArXiv:2505.19988 [cs].

Shayan Talaei, Mohammadreza Pourreza, Yu-Chen Chang, Azalia Mirhoseini, and Amin Saberi. 2024. CHESS: Contextual Harnessing for Efficient SQL Synthesis. *arXiv preprint*. ArXiv:2405.16755 [cs].

Qwen Team. 2025. Qwen3-next-80b-a3b: Qwen3-next: Towards ultimate training & inference efficiency. `https://qwen.ai/blog?id=qwen3-next`. Accessed: 2025-11-09.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2020. RAT-SQL: Relation-aware schema encoding and linking for text-to-SQL parsers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7567–7578, Online. Association for Computational Linguistics.

Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, LinZheng Chai, Zhao Yan, Qian-Wen Zhang, Di Yin, Xing Sun, and Zhoujun Li. 2025a. MAC-SQL: A multi-agent collaborative framework for text-to-SQL. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 540–557, Abu Dhabi, UAE. Association for Computational Linguistics.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schuetze. 2025b. Language mixing in reasoning language models: Patterns, impact, and internal causes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2637–2665, Suzhou, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhewei Yao, Guoheng Sun, Lukasz Borchmann, Zheyu Shen, Minghang Deng, Bohan Zhai, Hao Zhang, Ang Li, and Yuxiong He. 2025. Arctic-Text2SQL-R1: Simple Rewards, Strong Reasoning in Text-to-SQL. *arXiv preprint*. ArXiv:2505.20315 [cs].

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *Preprint*, arXiv:1709.00103.