Text-to-SQL na Era dos Grandes Modelos de Linguagem: Fundamentos, Avaliação e o Cenário Brasileiro

Guilherme Fonseca[1], Julio C. S. Reis[2], Matheus Botelho[1], Eduardo Camara[3], Jaide Zardin[3], Lucas R. Silva[1], Emerson A. Simoes[1], Pedro Calais[1], Rodrigo Gonçalves[1], Gustavo Ribeiro[1], Yago Lobato[3], Allan Sene[4], Altigran da Silva[3], Marcos André Gonçalves[1]

[1] Universidade Federal de Minas Gerais (UFMG),
[2] Universidade Federal de Viçosa (UFV),
[3] Universidade Federal do Amazonas (UFAM),
[4] Dadosfera Brasil

{guilhermefonseca, matheusbotelho, emerson.simoes}@dcc.ufmg.br
{pedrolgcalais, lucasrsilvak, rodrigopfmg, gustavo9661}@ufmg.br
{eduardo.camara, jaide.zardin, yagobrlobato, alti}@icomp.ufam.edu.br
jreis@ufv.br, allan@dadosfera.ai, mgoncalv@dcc.ufmg.br

Abstract. *The Text-to-SQL task has advanced significantly with the use of Large Language Models (LLMs). Despite this progress, evaluations conducted on traditional benchmarks such as Spider and BIRD may overestimate models' real-world capabilities, as evidenced by recent studies in enterprise settings and in Portuguese. This advanced tutorial systematically presents the foundations and state of the art of the task, covering: (i) the main approaches based on LLMs and agentic systems; (ii) a critical taxonomy of evaluation metrics; (iii) benchmark construction methodologies; and (iv) the current landscape of Text-to-SQL in Portuguese, with an emphasis on the Brazilian context.*
Keywords: Text-to-SQL; LLMs; evaluation metrics; benchmarks.

Resumo. *A tarefa Text-to-SQL tem avançado significativamente com o uso de Grandes Modelos de Linguagem (LLMs). Apesar desse progresso, avaliações conduzidas em benchmarks tradicionais como Spider e BIRD podem superestimar a capacidade real dos modelos, conforme evidenciado por trabalhos recentes em cenários empresariais e em língua portuguesa. Este tutorial avançado apresenta, de forma sistemática, os fundamentos e o estado da arte da tarefa, cobrindo: (i) as principais abordagens baseadas em LLMs e sistemas agênticos; (ii) uma taxonomia crítica de métricas de avaliação; (iii) metodologias de construção de benchmarks; e (iv) o panorama da tarefa em português, com ênfase no contexto brasileiro.*
Palavras-chave: Text-to-SQL; LLMs; métricas de avaliação; benchmarks.

1. Introdução

Nos últimos anos, o volume de informações produzidas e armazenadas tem crescido de forma exponencial, impulsionado, entre outros fatores, pela digitalização de processos governamentais e empresariais, o aumento do uso de sistemas transacionais e a expansão e barateamento da computação em nuvem. Em paralelo com esse aumento

cresce também a necessidade de realizar análises capazes de extrair informações relevantes desses dados. Tradicionalmente, esse processo é restrito a uma parcela específica de usuários com conhecimento em linguagens de consulta estruturadas, como o SQL, o que cria uma barreira de acesso significativa para tomadores de decisão, gestores e usuários finais sem formação técnica especializada, que, geralmente, são os principais interessados em consumir esse conhecimento. Nesse contexto, a tarefa *Text-to-SQL*, que consiste em traduzir perguntas formuladas em linguagem natural para consultas SQL executáveis, surge como uma solução para democratizar o acesso à informação estruturada [Katsogiannis-Meimarakis and Koutrika 2023]. Recentemente, com a chegada das arquiteturas *Transformers*, em particular dos LLMs (do inglês, *Large Language Models*) a tarefa tem tido avanços significativos [Yu et al. 2018, Gao et al. 2024, Shi et al. 2025].

Apesar desse progresso, a avaliação desses sistemas permanece fortemente concentrada no idioma inglês e em *benchmarks* isolados como Spider [Yu et al. 2018] e BIRD [Li et al. 2023], nos quais modelos recentes podem chegar a ultrapassar 80% de acurácia de execução [Gao et al. 2024, Shi et al. 2025]. Entretanto, trabalhos recentes mostram que esse desempenho não se transfere diretamente para cenários reais, o Spider2.0 [Lei et al. 2025], construído com *workflows* empresariais reais, mostrou que mesmo o estado da arte resolve apenas 21,3% das tarefas, contra 91,2% no Spider1.0, mostrando que *benchmarks* tradicionais podem superestimar sistematicamente a capacidade dos modelos. Padrão semelhante é reportado em *benchmarks* empresariais como o BEAVER [Chen et al. 2024], construído a partir de logs reais de queries privadas.

No contexto da língua portuguesa, embora existam iniciativas isoladas voltadas a avaliar a efetividade de modelos e as particularidades da tarefa, nota-se uma intensificação dos desafios metodológicos, com a maioria dos trabalhos se concentrando em bases isoladas, domínios específicos ou traduções de *benchmarks* em Inglês [de Carvalho Fróes and Braghetto 2025, Yamate et al. 2025, Jose and Cozman 2023, Pedroso et al. 2025]. Esses trabalhos, também, adotam protocolos experimentais diferentes, variando modelos avaliados, estratégias de *prompting* e métricas reportadas, o que dificulta comparações diretas entre estudos e compromete a avaliação realista do estado da arte.

Neste contexto, este tutorial aborda três dimensões da tarefa *Text-to-SQL*: (i) uma visão geral da tarefa, incluindo suas principais bases de dados, métodos e métricas de avaliação; (ii) diferenças entre *benchmarks* Tradicionais e Cenários Reais mais uma discussão sobre a construção de *benchmarks*; e (iii) um panorama sistematizado do ecossistema brasileiro de *Text-to-SQL*, com análise comparativa dos conjuntos de dados e resultados disponíveis na literatura.

2. Sumário do Tutorial e Descrição dos Tópicos

Evolução do *Text-to-SQL*. Ao longo da última década, a tarefa de *Text-to-SQL* sofreu transformações significativas, saindo de métodos baseados em regras e templates para sistemas baseados em aprendizado profundo e, atualmente, abordagens centradas em LLMs [Katsogiannis-Meimarakis and Koutrika 2023, Shi et al. 2025]. As primeiras tentativas na área eram baseadas em modelos probabilísticos acoplados a gramáticas formais e técnicas de correspondência de padrões. Embora funcionassem bem em cenários controlados, essas abordagens se mostravam bastante limitadas quando confrontadas com bancos de dados complexos e dinâmicos, ou com consultas que envolviam múltiplos domínios. Isso muda

com o avanço do aprendizado profundo, onde modelos *sequence-to-sequence* (Seq2Seq) supervisionados passaram a dominar a literatura [Zhong et al. 2017], sendo capazes de alinhar linguagem natural à estrutura física das tabelas por meio de codificadores de grafos e mecanismos de atenção, ao mesmo tempo que surgiram *benchmarks* de larga escala como WikiSQL e o Spider. Atualmente, modelos LLM, como GPT e LLaMA, pré-treinados em grandes volumes de texto e código, se apresentam como o estado da arte, sendo capazes de superar modelos supervisionados específicos desenvolvidos anteriormente [Shi et al. 2025]. Este módulo abordará essa trajetória evolutiva, discutindo vantagens, limitações e implicações dessas diferentes gerações de métodos de *Text-to-SQL*.

Abordagens Modernas. Apesar de LLMs serem o paradigma dominante, abordagens puramente *zero-shot* frequentemente se mostram insuficientes para cenários complexos, e para contornar isso a literatura vem focando em técnicas para melhorar o desempenho dos modelos. Uma delas é melhorar as estratégias de *prompting*, introduzindo raciocínio por meio de *Chain-of-Thought* (CoT) e estratégias *few-shot* baseadas na seleção de exemplos que auxiliam o modelo na hora da resposta. Outros métodos focam em melhorar o contexto dado ao modelo, isso por meio de Geração Aumentada por Recuperação (RAG), modelagem de *views* compatíveis com LLMs ou *schema linking* que mitiga alucinações em esquemas com dezenas de tabelas. Outro eixo é introduzir aos métodos um design agêntico, nos quais sistemas como MAC-SQL [Wang et al. 2025] decompõem a tarefa em etapas especializadas, incluindo identificação de escopo, planejamento, geração e correção iterativa pós-execução. Este módulo detalhará as principais técnicas e componentes arquiteturais que estruturam os sistemas modernos de *Text-to-SQL*. Além disso, nesse módulo, iremos apresentar implementações práticas de modelos a serem disponibilizados à audiência.

Métricas de Avaliação. Por se tratar de uma tarefa de geração multi token, na qual múltiplas consultas SQL distintas podem responder corretamente a mesma pergunta em linguagem natural, a etapa de avaliação em *Text-to-SQL* constitui um passo muito importante, pois a escolha da métrica possui implicações diretas sobre as conclusões extraídas nos experimentos. Atualmente as métricas podem ser divididas em 2 grandes grupos. (i) métricas estruturais e sintáticas, que focam em comparar as estruturas dos SQLs em si, como o *Component Match*, *Levenshtein Correctness*, *Structural Correctness* e *Skeleton Correctness*; (ii) métricas baseadas em execução, que comparam os dados retornados pelos SQLs, como Acurácia de Execução – EX (métrica mais explorada em trabalhos neste contexto), *Valid Efficiency Score* e *Soft Execution F1* [Yu et al. 2018]. Este módulo apresentará as principais métricas utilizadas na avaliação de sistemas *Text-to-SQL*, discutindo seus pontos fortes e limitações, além de ilustrar, por meio de exemplos reais, como diferentes métricas podem influenciar a interpretação das saídas produzidas pelos métodos.

***Benchmarks* Tradicionais e Cenários Reais.** Outra decisão importante é a escolha dos *benchmarks* nos quais serão testados os métodos de *Text-to-SQL*. Este módulo promoverá uma análise comparativa e crítica entre as bases utilizadas na literatura e as demandas impostas por aplicações em ambientes de produção. Será discutido como *benchmarks* como o Spider [Yu et al. 2018] e BIRD [Li et al. 2023] se estabeleceram como as principais bases de dados utilizadas em artigos científicos, mas tem comportamentos e características distintas das encontradas em cenários reais de empresas ou indústrias. Em contrapartida, discutiremos o declínio expressivo de desempenho dos modelos quando aplicados a bases de dados mais complexas e que se assemelham mais a esses cenários,

como Spider 2.0 [Lei et al. 2025] e BEAVER [Chen et al. 2024]. Faremos, também, uma breve discussão sobre a criação de *benchmarks*, apresentando pipelines de geração baseados em LLMs como o CodeS [Li et al. 2024], BenchPress [Wenz et al. 2025] e PAIRS [Camara et al. 2026] e discutindo suas limitações.

***Text-to-SQL* em Português.** Em língua portuguesa, a tarefa de *Text-to-SQL* também vem ganhando força nos últimos anos com o aumento do número de trabalhos. Porém, os estudos se mostram sempre iniciativas concentradas em domínios específicos e protocolos experimentais diversos, com cada estudo utilizando bases de dados próprias e modelos distintos. Este módulo apresentará um panorama dos esforços e recursos existentes para a tarefa em português, cobrindo as traduções do Spider [Jose and Cozman 2023, Pedroso et al. 2025] e os *benchmarks* específicos da língua portuguesa, como text2SQL4PM [Yamate et al. 2025], DATASUS [Moraes et al. 2025] e JABUTI-SQL [Botelho et al. 2026]. Discutiremos ainda nosso trabalho recente [Fonseca et al. 2026], que consolida essas quatro bases sob um protocolo experimental unificado. A análise integrada dos trabalhos discutidos nesse e nos módulos anteriores permitirá identificar padrões que orientam o avanço da área.

Uso de Inteligência Artificial. Utilizamos os modelos Claude e Gemini para revisão textual, com supervisão e revisão dos autores.

Agradecimentos. INCT-TILD-IAR (#408490/2024-1), FAPEMIG, CAPES, Dadosfera, Embrapii, SEBRAE.

Referências

Botelho, M. et al. (2026). Jabuti-sql: Um benchmark em português para avaliação de abordagens text-to-sql em cenários reais. In *DSW/SBBD*.

Camara, E., Zardin, J., et al. (2026). Pairs: Um pipeline para geração automática e curadoria de benchmarks text-to-sql. In *SBBD*.

Chen, P. B., Yang, D., Li, W., Wenz, F., Zhang, Y., Tatbul, N., Cafarella, M., Demiralp, Ç., and Stonebraker, M. (2024). Beaver: an enterprise benchmark for text-to-sql. *arXiv preprint arXiv:2409.02038*.

de Carvalho Fróes, K. and Braghetto, K. R. (2025). Exploring temporal text-to-sql challenges in brazilian portuguese: Lessons from educational data. In *SBBD*, pages 963–969.

Fonseca, G. et al. (2026). Avaliação sistemática de text-to-sql em português: Um benchmark unificado multi-domínio. In *SBBD*.

Gao, D. et al. (2024). Text-to-sql empowered by large language models: A benchmark evaluation. *VLDB Endowment*, 17(5):1132–1145.

Jose, M. A. and Cozman, F. G. (2023). A multilingual translator to sql with database schema pruning to improve self-attention. *Int'l Journal of Information Technology*, 15(6):3015–3023.

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

Lei, F. et al. (2025). Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. In *ICLR*, pages 28691–28735.

Li, H. et al. (2024). Codes: Towards building open-source language models for text-to-sql. *PACMMOD*.

Li, J. et al. (2023). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *NeurIPS*, 36:42330–42357.

Moraes, M. et al. (2025). From questions to answers: A natural language interface for datasus hospitalization data. In *ERAMIA-RS*, pages 296–299.

Pedroso, B. C., Pereira, M. R., and Pereira, D. A. (2025). Performance evaluation of llms in the text-to-sql task in portuguese. In *SBSI*, pages 260–269.

Shi, L. et al. (2025). A survey on employing large language models for text-to-sql tasks. *ACM Computing Surveys*, 58(2):1–37.

Wang, B. et al. (2025). Mac-sql: A multi-agent collaborative framework for text-to-sql. In *COLING*.

Wenz, F. et al. (2025). Benchpress: A human-in-the-loop annotation system for rapid text-to-sql benchmark curation. *arXiv preprint arXiv:2510.13853*.

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. et al. (2018). Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *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*.