PAIRS: Um *Pipeline* para Geração Automática e Curadoria de *Benchmarks Text-to-SQL*

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

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

eduardo.camara@icomp.ufam.edu.br, jaide.zardin@icomp.ufam.edu.br

Abstract. *This work presents the* PAIRS *(**P**ersona-**A**ware **I**ntents for **R**eal-world **SQL**), a pipeline for the automatic generation and curation of Natural Language–SQL pairs with Human-in-the-Loop support, accessible through a Web interface. The approach integrates automatic generation and validation of SQL queries, synthesis of questions conditioned on different personas, paraphrasing techniques to increase linguistic diversity, and human curation in a unified workflow. By simulating different user profiles and data literacy levels, the pipeline aims to produce semantically consistent Text-to-SQL benchmarks that better reflect real-world usage scenarios. As a case study, the approach was evaluated in the public health domain using DataSUS databases. The results demonstrate the pipeline's potential to support the interactive construction of specialized benchmarks and improve the evaluation of Text-to-SQL systems. The* PAIRS *implementation is open-source and available in a public repository, along with a demonstration video of the approach* [1].

Resumo. *Este trabalho apresenta o* PAIRS *(**P**ersona-**A**ware **I**ntents for **R**eal-world **SQL**), um pipeline para geração automática e curadoria de pares Linguagem Natural–SQL com suporte Human-in-the-Loop, disponibilizado por meio de uma interface Web. A abordagem integra geração e validação automática de consultas SQL, síntese de perguntas condicionadas a diferentes personas, técnicas de paráfrase para ampliação da diversidade linguística e curadoria humana em um fluxo unificado. Ao simular diferentes perfis de usuários e níveis de letramento em dados, o pipeline busca produzir benchmarks Text-to-SQL semanticamente consistentes e mais próximos de cenários reais de uso. Como estudo de caso, a abordagem foi avaliada no domínio da saúde pública utilizando bases do DataSUS. Os resultados demonstram o potencial do* PAIRS *para apoiar a construção interativa de benchmarks especializados e aprimorar a avaliação de sistemas Text-to-SQL. A implementação do* PAIRS *possui código aberto disponível em repositório público, juntamente com um vídeo demonstrativo da abordagem*[1].

*Ambos os autores contribuíram igualmente para realização deste trabalho.

[1](Repositório): https://github.com/Dadosfera-DCC-UFMG/PAIRS e (Vídeo): https://youtu.be/Sb-VwcFEew4

1. Introdução

Os avanços em Grandes Modelos de Linguagem (LLMs) impulsionaram o desenvolvimento de sistemas *Text-to-SQL* [Katsogiannis-Meimarakis and Koutrika 2023], tecnologias que buscam democratizar o acesso a dados por meio de consultas em linguagem natural. Para treinamento e avaliação desses modelos, a comunidade consolidou *benchmarks* públicos de larga escala, como Spider [Yu et al. 2018] e BIRD [Li et al. 2024b], compostos por pares de pergunta e consulta SQL correspondente.

Entretanto, estudos recentes indicam que esses *benchmarks* frequentemente não refletem a complexidade de bancos corporativos reais [Fonseca et al. 2026], caracterizados por esquemas extensos, ambiguidades semânticas e terminologias específicas de domínio [Coelho et al. 2024, Wenz et al. 2026]. Além disso, abordagens sintéticas baseadas em paráfrases [Li et al. 2024a] e tradução reversa [Khalifat 2025], amplamente exploradas como recurso computacional neste contexto, tendem a produzir perguntas linguisticamente homogêneas, negligenciando diferentes perfis de usuários e níveis de conhecimento sobre os dados.

Para endereçar essas limitações, este trabalho apresenta o PAIRS, um *pipeline* para geração automática e curadoria de *benchmarks Text-to-SQL* orientados a domínios específicos. Utilizando o *GPT-5 mini* como motor generativo, a abordagem parte da geração e validação automática de consultas SQL contra o esquema do banco e, posteriormente, sintetiza perguntas condicionadas a diferentes *personas*. O fluxo também integra geração de paráfrases e curadoria humana (*Human-in-the-Loop*) em um ambiente unificado. Diferentemente de abordagens focadas apenas em variações sintáticas, o PAIRS busca produzir pares semanticamente consistentes, linguisticamente diversos e mais próximos de cenários reais de uso.

O PAIRS foi implementado e disponibilizado em uma interface Web de código aberto para construção interativa de *benchmarks*. Como estudo de caso, aplicou-se a abordagem ao domínio de saúde pública utilizando dados do DataSUS, resultando em 195 pares validados, com 94,2% de aproveitamento e concordância moderada na curadoria ($\kappa = 0,50$). Os resultados demonstram o potencial da abordagem para acelerar a criação de *benchmarks* especializados e apoiar o desenvolvimento de sistemas *Text-to-SQL* em cenários reais.

2. Trabalhos Relacionados

A geração automática de *benchmarks* para *Text-to-SQL* busca reduzir o custo de curadoria manual e adaptar dados a novos domínios. Abordagens como o CodeS [Li et al. 2024a] utilizam LLMs para parafrasear instâncias ou preencher *templates*, mas produzem principalmente variações sintáticas. Outras estratégias geram perguntas técnicas a partir de colunas previamente selecionadas e posteriormente as convertem em SQL [Coelho et al. 2024]. Para aumentar o controle estrutural, empregam-se técnicas de NER [Dragusin et al. 2025] ou geração parametrizada de SQL seguida de *back-translation*, como no SynthSQL [Khalifat 2025]. Embora eficazes, esses métodos tendem a gerar perguntas linguisticamente homogêneas. Em contraste, abordagens com curadoria humana especializada, como text2SQL4PM [Yamate et al. 2025], oferecem maior naturalidade e diversidade, porém com baixa escalabilidade e alto custo.

Nesse contexto, o PAIRS combina geração automática de SQL validada diretamente no SGBD com síntese de perguntas condicionadas a diferentes *personas*, permitindo

Tabela 1. Características dos *frameworks* de geração de *benchmarks Text-to-SQL*.

Característica	[A]	[B]	[C]	[D]	[E]
Geração automática de NL	●	●	●	●	●
Geração automática de SQL	●	●	⊖	●	●
Ponto de partida centrado no SQL	○	●	○	●	●
Curadoria manual e correção	⊖	○	○	○	●
Geração de paráfrases	○	●	●	○	●
Controle de complexidade	⊖	○	○	●	●
Validação de qualidade / autocorreção	○	○	●	●	●
Vínculo explícito pergunta-colunas	●	○	●	○	⊖
Múltiplas perguntas por SQL	○	●	⊖	○	●
Perfis de usuário (*Personas*)	○	○	○	○	●

Legenda de Trabalhos: [A] Coelho *et al.* [Coelho et al. 2024]; **[B]** CodeS [Li et al. 2024a]; **[C]** Dragusin *et al.* [Dragusin et al. 2025]; **[D]** SynthSQL [Khalifat 2025]; **[E]** PAIRS - *Pipeline* proposto.

simular múltiplos perfis de usuários e níveis de letramento em dados. O fluxo também integra geração de paráfrases, autocorreção iterativa e curadoria *Human-in-the-Loop*, visando produzir pares semanticamente consistentes e linguisticamente mais diversos.

A Tabela 1 resume o posicionamento do PAIRS frente às abordagens analisadas, utilizando (●) para suporte completo, (⊖) para parcial e (○) para ausência da característica. O suporte parcial ao vínculo explícito entre perguntas e colunas decorre de uma decisão de projeto, uma vez que essa associação é estabelecida indiretamente via SQL e curadoria, e não de uma limitação estrutural do método.

3. Arquitetura do *Pipeline* Proposto

Esta seção descreve a arquitetura do PAIRS, um sistema para geração automática e curadoria de *benchmarks Text-to-SQL*. A abordagem integra geração de consultas SQL, síntese de perguntas em linguagem natural e curadoria *Human-in-the-Loop* em um fluxo orientado a domínios específicos. A partir do esquema relacional (extraído diretamente do banco de dados), metadados semânticos e *personas*, o processo opera nas três etapas interconectadas a seguir (Figura 1) para produzir pares variados e semanticamente consistentes.

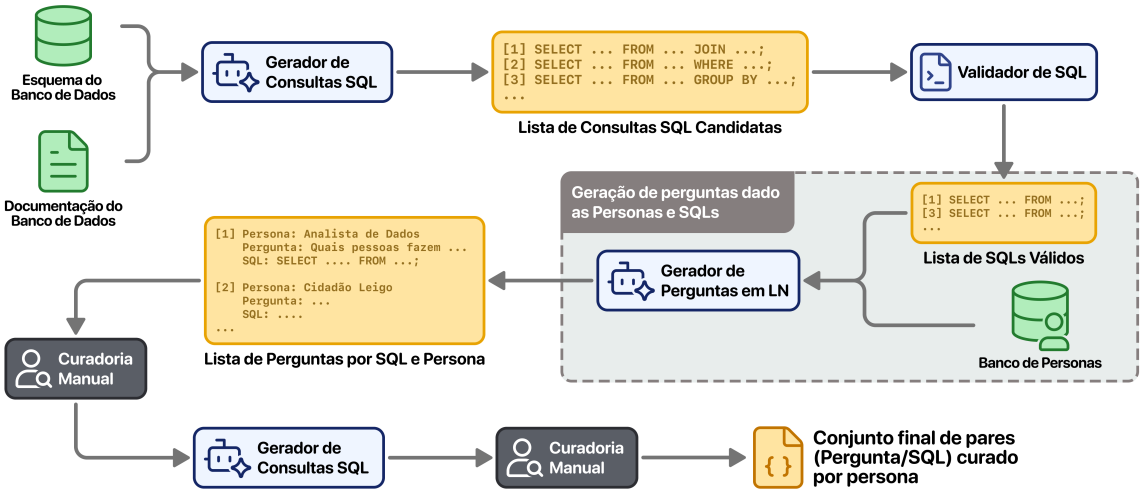

Figura 1. *Pipeline* de geração e curadoria *Human-in-the-Loop* do PAIRS.

Geração e Validação Automática de SQL. Inspirado no SynthSQL [Khalifat 2025], o sistema utiliza o *GPT-5 mini* para gerar consultas SQL a partir da definição do esquema rela-

cional, dos metadados e de níveis de complexidade (simples, médio e difícil), dispensando perguntas prévias em linguagem natural. As consultas passam por validação adaptativa — execução real quando há acesso ao SGBD via conector `psycopg2` ou análise estática baseada nas restrições estruturais do esquema utilizando a ferramenta `sqlglot` — e erros acionam um mecanismo iterativo de autocorreção por até N tentativas ($N = 3$ nos experimentos). Consultas inválidas ou com resultados vazios são descartadas. Embora os experimentos utilizem o *GPT-5 mini*, o PAIRS possui arquitetura agnóstica ao modelo, permitindo o uso de qualquer LLM; contudo, a versão atual suporta exclusivamente a API da OpenAI.

Geração Contextual Baseada em *Personas*. Com os SQLs validados, o sistema sintetiza perguntas condicionadas a diferentes *personas*, representando distintos perfis de usuários previamente definidos. Para cada par (*persona*, SQL), o modelo gera perguntas alinhadas ao contexto e vocabulário esperados. A Tabela 2 apresenta exemplos de saída desta etapa.

Tabela 2. Exemplos de perguntas geradas para diferentes personas e suas respectivas consultas SQL.

Persona	Pergunta	SQL Gerado
Cidadã Leiga	Quais são os 50 fornecedores com maior gasto em 2021...	`SELECT fornecedor_id, COUNT(*), SUM(preco_total) ... LIMIT 50;`
Analista	Para cada fornecedor, calcule o número de compras e total gasto em 2021...	
Cidadã Leiga	Quantas instituições tinham pelo menos um leito de UTI em 31/12/2022?	`SELECT COUNT(DISTINCT instituicao_id) ... WHERE ... > 0;`
Analista	Contagem de instituições distintas que reportaram leitos de UTI...	

Aumento de Dados com Paráfrases. Para ampliar a diversidade do *benchmark*, aplica-se um aumento de dados inspirado no CodeS [Li et al. 2024a]. O LLM gera paráfrases alterando a estrutura sintática, as entidades e as restrições (ex: "TOP 10" para "TOP 3"). Como essas alterações textuais podem invalidar o par original, o sistema reajusta automaticamente o SQL correspondente utilizando a biblioteca `sqlglot` para o *parsing* e manipulação direta da Árvore de Sintaxe Abstrata (AST) da consulta, modificando apenas as cláusulas necessárias. Por fim, uma última etapa de curadoria humana (*Human-in-the-Loop*) tem como objetivo corrigir inconsistências remanescentes e validar o alinhamento semântico.

4. Interface Web, Disponibilização e Utilização

O PAIRS possui código aberto em repositório público e está disponível em uma interface Web que viabiliza a construção interativa de *benchmarks Text-to-SQL*[1], integrando geração automática, validação e curadoria humana em um fluxo unificado.

A interface Web organiza o processo em cinco etapas: (1) carregamento da definição do esquema relacional, documentação e conexão com o SGBD; (2) geração e validação automática de consultas SQL; (3) revisão e filtragem das consultas geradas; (4) configuração de *personas* e síntese das perguntas; e (5) geração de paráfrases e curadoria *Human-in-the-Loop*. A etapa final disponibiliza um painel interativo para inspeção, edição, aprovação ou descarte dos pares antes da exportação do *benchmark*. A tela inicial da interface é apresentada na Figura 2.

O sistema foi construído com `streamlit` e integra diretamente o pacote do PAIRS disponibilizado no repositório. A interface aceita o arquivo de definição do banco

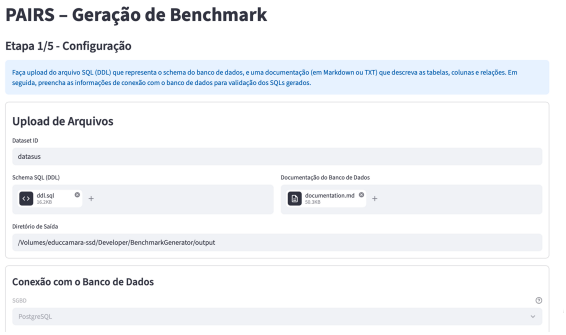

Figura 2. Página inicial do PAIRS.

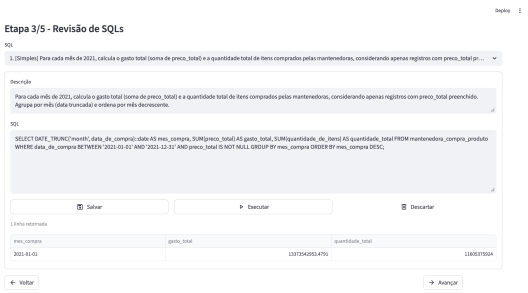

Figura 3. Interface da Etapa 3.

em formato `sql`, documentação em `txt` ou `md`, e exporta os pares gerados em formato `json`. A Figura 3 ilustra a etapa de revisão manual das consultas SQL geradas antes da síntese das perguntas.

5. Avaliação da Abordagem Proposta

Para avaliar a eficácia técnica e a variabilidade linguística do PAIRS, realizou-se um experimento utilizando uma base relacional real do domínio da saúde pública brasileira (DataSUS), analisando escalabilidade da geração e qualidade semântica dos pares produzidos.

5.1. Execução do *Pipeline* e Geração de Dados

Inicialmente, o sistema gerou 30 consultas SQL distribuídas entre os níveis fácil, médio e difícil. Após validação automática, autocorreção e descarte de consultas inválidas, foram retidas 23 consultas válidas contra o SGBD (10 fáceis, 6 médias e 7 difíceis).

Em seguida, aplicaram-se três *personas* (*Cidadã Leiga*, *Analista de Dados* e *Médico Gestor Hospitalar*), resultando em 69 pares (Pergunta, SQL). Posteriormente, o sistema gerou duas paráfrases para cada pergunta, produzindo um *corpus* candidato de 207 pares após nova validação automática.

5.2. Curadoria Humana (*Human-in-the-Loop*)

Três avaliadores independentes analisaram os pares quanto ao alinhamento entre pergunta, SQL e *persona*, classificando-os como aprovados, aceitos com ajustes leves ou reprovados. A confiabilidade do protocolo foi medida com o Coeficiente Kappa de Fleiss, obtendo $\kappa = 0,50$, correspondente a concordância moderada. As inconsistências nas análises foram mais frequentes nos itens de maiores níveis de dificuldade (médios e difíceis). Embora aceitável dada a subjetividade da tarefa, o resultado indica necessidade de maior padronização nos critérios de inspeção.

Após consenso e correções das amostras intermediárias, o *benchmark* final resultou em 195 pares validados, correspondendo a uma taxa de aproveitamento de 94,2%. Para avaliar as paráfrases, utilizaram-se métricas de similaridade semântica (SBERT) e sintática (BLEU). Os resultados ($BLEU = 0,50 \pm 0,13$ e $SBERT = 0,92 \pm 0,03$) indicam diversidade linguística com preservação da consistência semântica.

6. Conclusão e Trabalhos Futuros

O PAIRS demonstrou viabilidade técnica para apoiar a criação de *benchmarks Text-to-SQL*, combinando geração automática de SQL, *personas*, paráfrases e curadoria *Human-in-the-*

Loop. A abordagem produziu pares linguisticamente diversos com garantia de execução. Entretanto, o estudo apresenta limitações relacionadas à escala reduzida do experimento de avaliação, ao uso de um único domínio e ao *trade-off* entre diversidade linguística e número de SQLs distintos (195 perguntas derivadas de 23 SQLs únicos). Além disso, o coeficiente Kappa moderado indica que a avaliação da adequação às *personas* ainda envolve subjetividade humana , necessitando de uma padronização na avaliação.

Como trabalhos futuros, pretende-se avaliar o PAIRS em múltiplos domínios, explorar estratégias *LLM-as-a-Judge* para escalonar a avaliação semântica e ampliar a diversificação automática de intenções e metadados associados à estrutura do banco de dados.

Uso de Inteligência Artificial

Os modelos Claude e Gemini foram utilizados para revisão textual, tradução e tratamento dos dados, sob supervisão e revisão dos autores.

Agradecimentos

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

Referências

Coelho, G. M. C., Nascimento, E. R. S., Izquierdo, Y. T., García, G. M., Feijó, L., Lemos, M., Garcia, R. L. S., de Oliveira, A. R., Pinheiro, J. P., and Casanova, M. A. (2024). Improving the accuracy of text-to-sql tools based on large language models for real-world relational databases. In *Database and Expert Systems Applications*, pages 93–107.

Dragusin, C., Mirylenka, K., Czasch, C. M., Glass, M., Defosse, N., Scotton, P., and Gschwind, T. (2025). Grounding llms for database exploration: Intent scoping and paraphrasing for robust nl2sql. *Proceedings of the VLDB Endowment*, page 8097.

Fonseca, G., Reis, J. C. S., Botelho, M., Calais, P., Simoes, E. A., Silva, L. R., Camara, E., Zardin, J., Gonçalves, R., Ribeiro, G., Lobato, Y., Sene, A., da Silva, A., and Gonçalves, M. A. (2026). Avaliação sistemática de text-to-sql em português: Um benchmark unificado multi-domínio. In *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*.

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

Khalifat, N. (2025). Synthsql: A framework for generating synthetic text-to-sql datasets with controlled complexity. Technical report, University of Alberta.

Li, H., Zhang, J., Liu, H., Fan, J., Zhang, X., Zhu, J., Wei, R., Pan, H., Li, C., and Chen, H. (2024a). Codes: Towards building open-source language models for text-to-sql. *Proc. ACM Manag. Data*, 2(3).

Li, J., Hui, B., Qu, G., Yang, J., Li, B., et al. (2024b). Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *NeuriPS*, 36.

Wenz, F., Bouattour, O., Yang, D., Choi, J., Gregg, C., Tatbul, N., and Demiralp, Ç. (2026). Benchpress: A human-in-the-loop annotation system for rapid text-to-sql benchmark curation. In *CIDR*.

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

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., Zhang, Z., and Radev, D. (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.