Do Acesso à Resposta: Um Dataset Text-to-SQL em Português para Monitoramento de Políticas Públicas Educacionais*

Karina de Carvalho Fróes[1], **Kelly Rosa Braghetto**[1]

[1]Instituto de Matemática, Estatística e Ciência da Computação
Universidade de São Paulo

{karina.froes,kellyrb}@ime.usp.br

Abstract. *Text-to-SQL models enable querying databases through natural language, broadening access to public data. However, widely used benchmarks focus on generic domains and provide limited coverage for common public policy questions. This work presents a Portuguese dataset containing 107 question–SQL pairs built from the Brazilian School Census microdata. The questions address socially relevant topics, such as school meals, accessibility, infrastructure, and educational inequalities, covering different levels of difficulty and SQL complexity. This work also discusses challenges in natural language interfaces and proposes guidelines for building similar resources in other public domains.*

Resumo. *Modelos Text-to-SQL permitem consultar bases de dados por linguagem natural, ampliando o acesso a dados públicos. Entretanto, benchmarks amplamente utilizados concentram-se em domínios genéricos e oferecem cobertura limitada para perguntas comuns em políticas públicas. Este trabalho apresenta um dataset em português com 107 pares pergunta–SQL construídos a partir dos microdados do Censo Escolar Brasileiro. As perguntas abordam temas de impacto social, como alimentação, acessibilidade, infraestrutura e desigualdades educacionais, com diferentes níveis de dificuldade e complexidade SQL. O trabalho também discute desafios nas interfaces de linguagem natural e propõe diretrizes para recursos semelhantes em outros domínios públicos.*

1. Introdução

A crescente digitalização das atividades econômicas, sociais e governamentais tem resultado em um aumento expressivo na produção e disponibilização de dados. No setor público, esse movimento impulsionou iniciativas de dados abertos em diversos países, com o objetivo de promover transparência, estimular inovação e ampliar a participação social [Attard et al. 2015]. Exemplos incluem o portal Data.gov [Data.gov 2009] nos Estados Unidos, a *Open Government Partnership* [Open Government Partnership 2011] em âmbito internacional e, no Brasil, diferentes iniciativas de disponibilização de dados governamentais por órgãos públicos. Entre as bases brasileiras, destacam-se os microdados do Censo Escolar da Educação Básica, divulgados anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), que oferecem um panorama detalhado sobre escolas, matrículas, docentes e infraestrutura educacional em todo o país [INEP 2026].

*Funding: This work was supported by São Paulo Research Foundation (FAPESP) [grants #2023/00779-0 and #2023/18026-8]; and the National Council for Scientific and Technological Development (CNPq) [grant #420623/2023-0].

Esses dados permitem análises sobre temas de elevado interesse social, como acessibilidade, alimentação escolar, desigualdades raciais e de gênero, qualificação docente e infraestrutura básica, apoiando o monitoramento de políticas públicas e a formulação de intervenções baseadas em evidências. Entretanto, apesar de serem públicos, esses dados permanecem pouco acessíveis para grande parte da população, pois sua exploração geralmente exige domínio de SQL, compreensão de esquemas relacionais extensos e capacidade de análise de dados históricos.

Nesse contexto, interfaces em linguagem natural para bancos de dados (*Natural Language Interfaces to Databases* – NLIDBs), particularmente as abordagens Text-to-SQL, surgem como alternativa promissora para democratizar o acesso a dados estruturados [Visperas et al. 2023]. A ideia é permitir que usuários formulem perguntas como "Quais escolas melhoraram os indicadores de acessibilidade no último ano?" e o sistema gere e execute automaticamente a consulta SQL equivalente, devolvendo apenas o resultado ao usuário. Avanços recentes em *Large Language Models* (LLMs) ampliaram significativamente o potencial dessas soluções [Sun et al. 2023], tornando mais plausível a geração automática de consultas complexas.

Apesar desse progresso, *benchmarks* que são amplamente utilizados para treinamento e aperfeiçoamento de modelos Text-to-SQL, como Spider [Yu and et al. 2018], WikiSQL [Zhong et al. 2017] e BIRD [Li et al. 2024], concentram-se predominantemente na língua inglesa e em domínios de propósito geral, não contemplando perguntas socialmente relevantes sobre o tema de políticas públicas. Em cenários reais, essas consultas frequentemente exigem raciocínio temporal, inferências implícitas, conceitos abstratos e a integração de múltiplas tabelas relacionais.

Para abordar essa limitação, este trabalho apresenta um *dataset* em português composto por 107 pares pergunta–SQL construídos a partir dos microdados do Censo Escolar da Educação Básica brasileiro, visando contribuir para o aperfeiçoamento dos modelos Text-to-SQL. As perguntas abordam temas sociais e contemplam consultas analíticas complexas que envolvem múltiplos *joins*, funções de janela, análises temporais, rankings e inferências implícitas. Mais do que disponibilizar um novo recurso, o trabalho apresenta diretrizes para a construção de *datasets* voltados a políticas públicas, contemplando diferentes perspectivas analíticas (Seção 4) e uma classificação de consultas SQL por níveis de dificuldade, do fácil ao muito difícil (Seção 5.2). Esse tipo de iniciativa pode servir de referência para subsidiar soluções baseadas em Inteligência Artificial, como *Retrieval-Augmented Generation* (RAG) e agentes conversacionais, contribuindo para ampliar a acessibilidade aos portais de dados governamentais.

O restante deste artigo está organizado da seguinte forma. A Seção 2 aborda o uso dos dados do Censo Escolar no monitoramento de políticas públicas. A Seção 3 discute os trabalhos relacionados. A Seção 4 descreve os desafios para interfaces em linguagem natural contemplados pelo *dataset*. A Seção 5 apresenta a construção e caracterização do recurso proposto. Por fim, a Seção 6 apresenta as conclusões e considerações finais.

2. Dados do Censo Escolar e o Plano Nacional de Educação (PNE)

O Brasil possui um dos sistemas mais abrangentes de coleta de dados educacionais do mundo: o Censo Escolar da Educação Básica, realizado anualmente pelo INEP. Sua base de microdados contém informações declaratórias de todas as escolas, matrículas, docentes

e turmas do país, permitindo análises longitudinais e transversais sobre a educação básica.

A relevância estratégica desses dados está diretamente associada ao novo Plano Nacional de Educação (PNE 2026–2036) [Brasil. Ministério da Educação 2026], encaminhado pelo Ministério da Educação e estruturado em 19 objetivos, 73 metas e 372 estratégias. Sua organização é composta por oito grandes temáticas, incluindo: educação infantil, alfabetização, ensino fundamental e médio, educação integral, diversidade e inclusão, educação profissional e tecnológica, educação superior e estrutura e funcionamento da educação básica.

O novo PNE constitui o principal instrumento de planejamento educacional de longo prazo do país e orienta a formulação, implementação e avaliação de políticas públicas em âmbito nacional, estadual e municipal. Entre seus objetivos, destacam-se a ampliação do acesso à educação infantil, a melhoria da aprendizagem, a expansão da educação em tempo integral, o fortalecimento da inclusão educacional, a valorização dos profissionais da educação e o aprimoramento da infraestrutura escolar.

Para que o PNE seja efetivamente monitorado, é necessário transformar seus objetivos e metas, expressos em linguagem de políticas públicas, em perguntas analíticas operacionalizáveis a partir de bases de dados. Uma das principais fontes de informação para esse acompanhamento é o Censo Escolar. A Tabela 1 apresenta exemplos ilustrativos dessa relação.

Tabela 1. Exemplos de perguntas analíticas associadas a objetivos do PNE

Objetivo do PNE	**Exemplo de pergunta analítica**
Aprendizagem no Ensino Fundamental e Médio	Como evoluiu a taxa de distorção idade-série por região nos últimos cinco anos?
Diversidade e Inclusão	Qual a proporção de estudantes negros, quilombolas e indígenas na educação básica por unidade da federação nos últimos três anos?
Profissionais da Educação Básica	Como evoluiu a proporção de docentes com formação adequada à etapa de ensino em que atuam nas redes estaduais ao longo da última década?
Financiamento e Infraestrutura da Educação	Quais são e onde estão localizadas as escolas rurais que ainda não dispõem de abastecimento de água, esgoto sanitário e acesso à internet?

Cada pergunta apresentada na Tabela 1, embora natural para gestores, pesquisadores ou jornalistas de dados, pode exigir consultas SQL complexas, envolvendo múltiplos *joins* entre tabelas de matrículas, escolas e docentes, cálculos de proporções, comparações históricas e inferências indiretas. Além disso, alguns conceitos presentes no PNE não correspondem diretamente a um único atributo da base de dados do Censo. Por exemplo, no caso dos profissionais da educação, a expressão "formação adequada" representa um conceito normativo e multidimensional, cuja operacionalização depende da combinação de diferentes variáveis do Censo Escolar, como nível de escolaridade, área de formação, componente curricular ministrado e etapa de ensino em que o docente atua. A resposta

à pergunta exige, portanto, a tradução de um conceito abstrato de política pública em regras analíticas explícitas, ilustrando a distância semântica entre a linguagem utilizada por especialistas do domínio e as estruturas efetivamente disponíveis na base de dados.

Assim, o PNE oferece um arcabouço de alto valor social para a formulação de perguntas relevantes e concretas sobre a educação brasileira. A capacidade de responder rapidamente a essas perguntas a partir dos dados pode apoiar o controle social, o planejamento governamental e a formulação de políticas públicas baseadas em evidências.

Embora os microdados do Censo Escolar sejam públicos e gratuitos, sua exploração permanece restrita a usuários com conhecimento técnico especializado. Essa dificuldade decorre principalmente de: (i) esquemas relacionais extensos e multirrelacionais; (ii) necessidade de raciocínio temporal para análises históricas e detecção de transições; (iii) conceitos inferenciais que não possuem representação direta na base; e (iv) coexistência de linguagem técnica educacional com expressões naturais utilizadas por gestores e cidadãos.

Como consequência, análises potencialmente relevantes para monitoramento de políticas públicas acabam concentradas em grupos tecnicamente especializados, limitando o alcance social das iniciativas de transparência governamental.

3. Trabalhos Relacionados

Iniciativas de *Data Science for Social Good* (DS4SG) frequentemente envolvem a construção, integração e organização de bases de dados capazes de apoiar análises sobre problemas sociais complexos. Sob essa ótica, trabalhos brasileiros como o InSANdB [Oliveira et al. 2025] e o *Brazil Data Commons* [Rodrigues et al. 2025] buscam estruturar e integrar dados públicos para ampliar seu potencial analítico, reutilização científica e apoio à formulação de políticas públicas por meio de *datasets* especializados.

Além da construção de *datasets*, iniciativas recentes também têm explorado técnicas de Ciência de Dados e Inteligência Artificial para ampliar a transparência pública e apoiar processos de auditoria governamental. O trabalho de análise de gastos públicos com dados do SICOM [Costa et al. 2024], por exemplo, descreve o uso de mineração de dados, processamento de linguagem natural e ferramentas analíticas para fiscalização de licitações municipais e detecção de irregularidades. Os autores destacam desafios recorrentes em bases governamentais reais, como baixa padronização, heterogeneidade estrutural e dificuldades de acesso e interpretação dos dados públicos.

Esses trabalhos evidenciam que iniciativas de DS4SG envolvem não apenas a disponibilização de dados, mas também a criação de recursos capazes de tornar as informações públicas mais acessíveis, interpretáveis e reutilizáveis em aplicações computacionais. Nesse contexto, as interfaces em linguagem natural representam uma alternativa promissora para democratizar o acesso a bases governamentais complexas.

No âmbito de *datasets* para Text-to-SQL, *benchmarks* como WikiSQL [Zhong et al. 2017], Spider [Yu and et al. 2018] e BIRD [Li et al. 2024] ampliaram progressivamente a complexidade estrutural e a capacidade de generalização entre esquemas. Entretanto, esses recursos concentram-se predominantemente no inglês e em domínios de propósito geral, oferecendo cobertura limitada para perguntas analíticas típicas de políticas públicas, frequentemente marcadas por raciocínio temporal,

inferências implícitas e conceitos semânticos abstratos.

Como evidência empírica dessa limitação, as autoras do presente trabalho conduziram uma análise do desempenho de uma *pipeline* de Text-to-SQL conhecida como DIN-SQL [Pourreza and Rafiei 2024], usando como modelo de linguagem o GPT-4, em um subconjunto de perguntas do *dataset* aqui trabalhado [Fróes and Braghetto 2025]. Os resultados revelaram limitações sistemáticas em aspectos centrais para o domínio de políticas públicas: (i) dificuldade em raciocínio temporal dinâmico; (ii) incapacidade de realizar inferências multietapa, como a detecção do fechamento de escolas a partir da ausência de registros entre censos consecutivos; (iii) alucinação de colunas inexistentes no esquema relacional; e (iv) falhas no alinhamento semântico entre a pergunta em português e as estruturas disponíveis no banco de dados. Essas observações reforçam a necessidade de *datasets* que representem adequadamente a complexidade analítica e temporal de perguntas reais sobre dados governamentais.

O presente trabalho busca mitigar essa lacuna ao introduzir um *dataset* em português focado em políticas públicas educacionais, projetado especificamente para atender a consultas analíticas complexas. O recurso contempla desafios de temporalidade, raciocínio em múltiplas etapas, inferência e múltiplas tabelas relacionais, complementando os *benchmarks* tradicionais ao aproximar a tarefa de Text-to-SQL dos cenários reais e dinâmicos da gestão pública.

4. Desafios para Interfaces em Linguagem Natural Abordados no *Dataset*

Esta seção descreve desafios importantes com os quais sistemas que fornecem interfaces em linguagem natural para acesso a bancos de dados precisam lidar. Esses desafios serviram como princípios orientadores para a construção do *dataset* descrito na Seção 5, garantindo a representação sistemática de fenômenos relevantes ao desenvolvimento de sistemas Text-to-SQL aplicados ao monitoramento de políticas públicas.

4.1. Ambiguidade Semântica

Ambiguidade semântica ocorre quando uma pergunta admite múltiplas interpretações válidas. Em dados públicos, isso inclui ambiguidades conceituais, de granularidade e referenciais. Termos como “melhorou”, “acesso” ou “qualidade” não possuem correspondência direta no esquema relacional e podem exigir interpretações distintas. Da mesma forma, perguntas como “Como evoluiu a proporção de docentes com doutorado?” não explicitam o nível de agregação desejado (escola, município, estado, país), enquanto entidades como “São Paulo” podem representar município ou estado.

Para ilustrar esse cenário, temos a seguinte pergunta no *dataset*: “Em quais escolas está concentrada a maioria das matrículas de alunos pretos em Itupeva em 2024?” O termo “concentrada” pode indicar maior número absoluto, maior proporção relativa ou conjunto acumulado acima de 50% do total municipal, produzindo consultas SQL distintas.

4.2. Variabilidade Linguística

Variabilidade linguística refere-se às múltiplas formas de expressar a mesma intenção. No domínio educacional, perguntas como “Quais escolas foram desativadas em 2023?”, “Mostre as escolas que fecharam em 2023” e “Quais entidades educacionais não aparecem mais no censo de 2023” devem produzir consultas SQL equivalentes. O mesmo ocorre

com expressões como “alunos negros”, “pretos e pardos” ou “distribuição por cor/raça”. Além de paráfrases sintáticas, o fenômeno envolve regionalismos e diferenças culturais, como o uso intercambiável de “merenda escolar” e “alimentação escolar”. O *dataset* incorpora explicitamente essas variações, permitindo avaliar robustez semântica.

4.3. Raciocínio Temporal e Tendências

Perguntas sobre políticas públicas frequentemente exigem análise temporal, que vai além de filtros simples por ano. O *dataset* contempla comparações históricas, detecção de transições de estado, médias móveis e análise de tendências. Perguntas como “Quais escolas passaram a oferecer alimentação escolar no último ano?” ou “A proporção de escolas com laboratório de informática apresentou crescimento ou declínio nos últimos seis anos?” exigem raciocínio temporal em múltiplas etapas, envolvendo funções de janela, comparação entre períodos e análise de séries temporais.

4.4. Inferência Lógica e Consultas Multi-Etapas

Muitas perguntas analíticas requerem múltiplas etapas de inferência e agregação. Um exemplo presente no *dataset* é: “Qual a taxa de evasão escolar por estado em 2024, considerando apenas escolas com pelo menos 100 alunos matriculados em 2023?” A consulta exige: (i) filtrar as escolas elegíveis com base no ano anterior; (ii) identificar, entre os alunos dessas escolas, aqueles que não se matricularam em 2024; e (iii) calcular a taxa proporcional agregada por estado, como mostrado na Figura 1. Essa sequência de filtro, identificação de ausência e agregação caracteriza o padrão de consulta multi-etapas presente em diversas perguntas do *dataset*.

Figura 1. Exemplo de Consulta SQL do *dataset*

```sql
-- Etapa 1: filtrar escolas elegiveis (>=100 alunos em 2023)
WITH escolas_elegiveis AS (
  SELECT id_escola, sigla_uf
  FROM matricula
  WHERE ano = 2023
  GROUP BY id_escola, sigla_uf
  HAVING COUNT(DISTINCT id_matricula) >= 100
),
-- Etapa 2: identificacao de evadidos (ausentes em 2024)
alunos_evadidos AS (
  SELECT e.sigla_uf, m23.id_matricula,
         CASE WHEN m24.id_matricula IS NULL THEN 1 ELSE 0 END AS evadiu
  FROM matricula m23
  JOIN escolas_elegiveis e ON m23.id_escola = e.id_escola
  LEFT JOIN matricula m24
    ON m23.id_matricula = m24.id_matricula AND m24.ano = 2024
  WHERE m23.ano = 2023
)
-- Etapa 3: calcular taxa agregada por estado
SELECT sigla_uf,
       SUM(evadiu) AS evadidos,
       COUNT(*) AS total_alunos,
       ROUND(SUM(evadiu) * 100.0 / COUNT(*), 2) AS taxa_evasao
FROM alunos_evadidos
GROUP BY sigla_uf;
```

4.5. Complexidade Estrutural dos Dados

O Censo Escolar possui uma estrutura relacional complexa, baseada em múltiplas tabelas com granularidades distintas, a saber: escola, matrícula, docente e turma, cada uma

com um elevado número de colunas. Isso introduz desafios relacionados tanto à seleção adequada de tabelas e colunas quanto à cardinalidade e à equivalência estrutural das consultas SQL. Exemplos como "Escolas com mais de 500 alunos e ao menos um professor com doutorado em 2020" exigem múltiplas junções e agregações adequadas. Diferentes estratégias SQL, como `NOT EXISTS` ou `LEFT JOIN`, também podem responder corretamente à mesma pergunta. No *dataset* criado, incluímos consultas que demandam integração de múltiplas tabelas e raciocínio estrutural complexo.

5. Construção e Caracterização do *Dataset*

O *dataset* foi construído manualmente a partir do conhecimento de domínio sobre o Censo Escolar e de demandas analíticas frequentemente associadas à formulação, ao monitoramento e à avaliação de políticas públicas, como é o caso do PNE. O processo envolveu: (i) seleção de temas de interesse social; (ii) elaboração de perguntas em linguagem natural em português; (iii) construção manual das consultas SQL correspondentes; e (iv) caracterização das perguntas quanto ao tema abordado e ao tipo de raciocínio necessário para sua resolução.

A elaboração manual permitiu incorporar cenários analíticos realistas, incluindo inferências implícitas, ambiguidades semânticas e análises históricas. Diferentemente de *benchmarks* genéricos, as perguntas foram concebidas para refletir demandas que poderiam ser formuladas por gestores públicos, pesquisadores, jornalistas de dados e cidadãos interessados em compreender a evolução de indicadores educacionais.

5.1. Caracterização Temática

O *dataset* reúne 107 pares pergunta–SQL sintaticamente válidas, cobrindo diferentes dimensões da educação básica brasileira. Entre os principais temas contemplados estão alimentação escolar, acessibilidade, mobilidade, saneamento, infraestrutura, qualificação docente, desigualdades de gênero e raça, evasão escolar e fechamento de unidades. A Tabela 2 apresenta a distribuição das perguntas por tema de política pública.

Tabela 2. Distribuição das perguntas por tema

Tema	Qtd.	%	Exemplos de subtemas
Matrículas	47	43,9%	gênero/raça/etnia, média móvel, evasão, vulnerabilidade, EJA, ranking
Docentes	22	20,6%	idade, titulação, formação, proporção aluno-professor
Infraestrutura	18	16,8%	merenda, biblioteca, internet, saneamento, água, energia, quadras
Acessibilidade	9	8,4%	PCD, ensino infantil, rampas, professores especializados, série histórica
Transição de Estado	7	6,5%	mudanças temporárias na oferta de serviços como alimentação, abertura/fechamento de escolas
Transporte	4	3,7%	meios de locomoção até a escola

Observa-se a predominância de perguntas de natureza longitudinal, evidenciando o foco do *dataset* em análises históricas e na avaliação da evolução de indicado-

res ao longo do tempo. Esse tipo de raciocínio é especialmente relevante em políticas públicas, nas quais o interesse frequentemente recai sobre tendências, transições e efeitos de intervenções governamentais.

5.2. Complexidade Analítica

Embora o objetivo principal deste trabalho não seja comparar modelos Text-to-SQL, o *dataset* foi projetado para representar consultas analíticas de diferentes níveis de dificuldade. As perguntas variam desde contagens simples até consultas que exigem comparações históricas, identificação de padrões temporais, médias móveis e inferências baseadas na ausência de registros.

Para caracterizar a complexidade das consultas, adotamos um sistema de pontuação multicritério que avalia nove dimensões estruturais como número de JOINs, uso de CTEs, *window functions*, subconsultas, operadores, agregação aninhada e quantidade de tabelas referenciadas. Cada dimensão contribui com uma pontuação específica, e a soma define o nível de complexidade: Fácil (0–2 pontos), Média (3–5), Difícil (6–9) ou Muito Difícil (10+). O detalhamento completo das regras de pontuação e o *dataset* estão disponíveis no repositório do GitHub [1]. Do total de consultas, 43% foram classificadas como Fácil, 28% como Média, e aproximadamente 29% como Difícil ou Muito Difícil. Esse último grupo indica que o *dataset* abrange desafios analíticos compatíveis com situações reais de exploração de dados públicos.

A Tabela 3 apresenta três perguntas representativas de diferentes níveis de complexidade, acompanhadas de suas respectivas taxas de *Execution Accuracy* (EX). O EX indica o percentual de consultas geradas que produzem exatamente o mesmo resultado da consulta *gold* (mesmas linhas e valores). Como esperado, os resultados mostram uma queda progressiva na precisão à medida que a complexidade da pergunta aumenta.

Tabela 3. Exemplos de perguntas do *dataset*, com nível de dificuldade e EX

Nível	Pergunta	EX
Fácil	Quantas escolas públicas de ensino médio ofereceram ensino em período integral no último ano?	13%
Média	Qual a proporção, por ano, de alunos por docente na rede pública e na rede privada nos últimos 5 anos?	10,3%
Muito Difícil	Existe alguma escola que não tinha alimentação escolar, passou a ter e manteve esse status nos anos seguintes? Considere 5 anos.	4,2%

Este trabalho destaca que muitas perguntas exigem raciocínio em múltiplas etapas, análise temporal e combinação de múltiplas fontes de informação. Esses aspectos tornam o *dataset* particularmente útil para o desenvolvimento de interfaces em linguagem natural voltadas ao acesso a dados governamentais.

5.3. Generalização para Outros Domínios

Embora tenha sido construído com dados educacionais, o *dataset* adota estruturas semânticas que podem ser reaproveitadas em outros domínios de interesse público. A

[1] https://github.com/froeska25/text-to-sql

seguir, apresentam-se exemplos reais do *dataset* e possíveis correspondências para o domínio de saúde pública:

- "Qual o total de escolas fechadas no último ano?" pode ser reformulada para verificar unidades de saúde desativadas;
- "Liste as 5 escolas públicas com maior proporção de alunos por professor nos últimos 5 anos para cada estado." pode ser reformulada para averiguar a média de pacientes por médico;
- "Houve aumento no número de professores das redes públicas com grau de escolaridade de mestre se compararmos o ano atual e o anterior?" pode ser reformulada para verificar a evolução da proporção de profissionais com especialização;
- "Quais escolas passaram a oferecer alimentação escolar em 2024?" pode ser reformulada para verificar quais municípios passaram a oferecer determinado serviço público de interesse.

Essa característica amplia o potencial de reutilização do *dataset* e reforça sua contribuição como arcabouço conceitual para o desenvolvimento de interfaces em linguagem natural em diferentes contextos de políticas públicas.

6. Conclusão

Este trabalho apresentou um *dataset* em português composto por perguntas em linguagem natural e suas respectivas consultas SQL, construído a partir de temas centrais para políticas públicas educacionais com base nos microdados do Censo Escolar Brasileiro. Além de seu diferencial de domínio, destaca-se também seu caráter técnico, contemplando perguntas analiticamente complexas que envolvem raciocínio temporal, inferências implícitas e integração entre múltiplas tabelas relacionais.

Mais do que um recurso técnico, este *dataset* representa um passo em direção à democratização do acesso a dados públicos e à transparência governamental. O recurso aproxima sistemas Text-to-SQL a cenários reais de análise de políticas públicas, alinhando-se à agenda de Ciência de Dados para o Bem Social. Embora desenvolvidos no domínio educacional, os padrões analíticos presentes no *dataset* podem ser adaptados a outros contextos de interesse público, como os de Saúde e Assistência Social.

Entretanto, ampliar o acesso a dados públicos por meio de interfaces em linguagem natural requer muito cuidado, pois também introduz desafios éticos importantes. Respostas produzidas automaticamente podem induzir interpretações equivocadas, simplificações excessivas ou inferências inadequadas, especialmente em temas relacionados a desigualdade social, raça e vulnerabilidade. Dessa forma, sistemas voltados ao acesso a dados governamentais devem incorporar mecanismos de transparência, explicabilidade e educação dos usuários sobre as limitações das respostas geradas.

Referências

Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418.

Brasil. Ministério da Educação (2026). Plano Nacional de Educação 2026–2036. `https://www2.camara.leg.br/legin/fed/lei/2026/lei-15388-14-abril-2026-798950-normaatualizada-pl.pdf`.

Documento preparatório do novo Plano Nacional de Educação (PNE 2026–2036). Acesso em: 20 maio 2026.

Costa, L., Dutra, M. T., Oliveira, G., Silva, M., Soares, D. C., de Cássia Silva Faria, L., Jr., W. M., and Pappa, G. (2024). Ciência de dados e transparência: Experiências com dados públicos do sicom. In *Anais Estendidos do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 246–252, Porto Alegre, RS, Brasil. SBC.

Data.gov (2009). The home of the U.S. Government's open data. `https://data.gov/`. Acesso em: 20 mai. 2026.

Fróes, K. d. C. and Braghetto, K. R. (2025). Exploring temporal Text-to-SQL challenges in brazilian portuguese: Lessons from educational data. *Anais do 40º Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 963–969.

INEP (2026). Microdados do Censo Escolar da Educação Básica. Disponível em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/censo-escolar`. Acesso em: 26 mai. 2026.

Li, J. et al. (2024). 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.

Oliveira, A. H., Martins, S., Rocha, D., and Nascimento, L. R. (2025). InSANdB: Base de dados integrada por município brasileiro para estudos da insegurança alimentar e nutricional no brasil. In *Simpósio Brasileiro de Sistemas de Informação*.

Open Government Partnership (2011). The open gov challenge. `https://www.opengovpartnership.org/`. Acesso em: 20 mai. 2026.

Pourreza, M. and Rafiei, D. (2024). Din-SQL: Decomposed in-context learning of text-to-SQL with self-correction. *Advances in Neural Information Processing Systems*, 36.

Rodrigues, I., Reis, J., Queiroz, B., and Benevenuto, F. (2025). BRAZIL DATA COMMONS: Plataforma unificada para análise de dados públicos de diferentes repositórios. In *Anais do XX Simpósio Brasileiro de Sistemas Colaborativos*, pages 308–316, Porto Alegre, RS, Brasil. SBC.

Sun, R. et al. (2023). SQL-PaLM: Improved large language model adaptation for Text-to-SQL (extended). *arXiv preprint arXiv:2306.00739*.

Visperas, M. et al. (2023). On modern Text-to-SQL semantic parsing methodologies for natural language interface to databases: A comparative study. In *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 390–396. IEEE.

Yu, T. and 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 the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

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