Universidade Federal de Pernambuco

Centro de Informática

Graduação em Sistemas de Informação

Uma Revisão Sistemática sobre Retrieval-Augmented Generation (RAG) e Modelos de Linguagem de Grande Escala (LLMs) na construção de consultas SQL

Trabalho de Conclusão de Curso de Graduação

por

Vinícius Marçal Araújo

Orientador: Prof. Robson Fidalgo

Recife, Fevereiro / 2025

Vinícius Marçal Araújo

Uma Revisão Sistemática sobre Retrieval-Augmented Generation (RAG) e Modelos de Linguagem de Grande Escala (LLMs) na construção de consultas SQL

Monografia apresentada em Sistemas de Informação, como requisito parcial para a obtenção do Título de Bacharel em Sistemas de Informação, Centro de Informática.

Orientador: Prof. Robson Fidalgo

Recife

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Araújo, Vinícius Marçal .
Uma Revisão Sistemática sobre Retrieval-Augmented Generation (RAG) e Modelos de Linguagem de Grande Escala (LLMs) na construção de consultas SQL / Vinícius Marçal Araújo. - Recife, 2025.
43 : il., tab.

Orientador(a): Robson do Nascimento Fidalgo
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Sistemas de Informação - Bacharelado, 2025.
Inclui referências.

1. Geração Aumentada de Recuperação. 2. Consultas SQL. 3. Modelos de Linguagem de Grande Escala. I. Fidalgo, Robson do Nascimento. (Orientação). II. Título.

600 CDD (22.ed.)

VINÍCIUS MARÇAL ARAÚJO

UMA REVISÃO SISTEMÁTICA SOBRE RETRIEVAL-AUGMENTED GENERATION (RAG) E MODELOS DE LINGUAGEM DE GRANDE ESCALA (LLMS) NA CONSTRUÇÃO DE CONSULTAS SQL

Trabalho de Conclusão de Curso (graduação) apresentada ao curso de Sistemas de Informação da Universidade Federal de Pernambuco, Centro de Informática, para obtenção do título de Bacharel em Sistemas de Informação.

.

Aprovado em: 27/02/2025

BANCA EXAMINADORA

_____
Prof. Dr. Robson Fidalgo (Orientadora)
Universidade Federal de Pernambuco - UFPE

_____
Prof. Dr. Luiz Augusto Morais (Examinador Interno)
Universidade Federal de Pernambuco - UFPE

Agradecimentos

Em primeiro lugar, quero agradecer à Doutora Fabiana, minha mãe, pois o maior exemplo de pesquisa e de estudante eu tive em casa, além de seu apoio incondicional, amor irrestrito e suporte para que eu pudesse virar quem eu sou, tanto em âmbito acadêmico como pessoal.

Agradeço também meu pai, Nivaldo, e a minha irmã, Maiara, que me ensinaram que o amor também serve para criticar, desde que seja para que eu seja sempre melhor.

Um agradecimento especial ao meu orientador, Robson, e a tantos outros professores do CIn e da UFPE que me acompanharam nessa jornada, que não seria possível sem eles.

Um agradecimento aos meus amigos, família que escolhi, que foram a base e o que possibilitou que eu chegasse até aqui como pessoa. Agradeço principalmente àqueles colegas que se transformaram em amigos, pois eles que sempre estiveram ao meu lado em toda minha graduação - sem essa parceria e amizade, minha passagem pela universidade seria no mínimo menos aproveitada por mim.

Por fim, um agradecimento a todos que acreditaram em mim e me apoiaram, e de que alguma maneira, também fizeram parte de minha trajetória e por isso merecem créditos nessa conquista.

"Computadores fazem arte,
artistas fazem dinheiro."

Fred Zero Quatro

RESUMO

Os Modelos de Linguagem de Grande Escala (LLMs) são modelos de inteligência artificial treinados em grandes quantidades de dados textuais, capazes de gerar respostas coerentes e realizar diversas tarefas de processamento de linguagem natural (PLN). A técnica de Retrieval-Augmented Generation (RAG) aprimora esses modelos ao incorporar fontes externas de conhecimento, permitindo que gerem respostas mais precisas e atualizadas. No contexto de bancos de dados, a conversão de linguagem natural para SQL (text-to-SQL) é um desafio, pois requer a interpretação correta da intenção do usuário e a geração de consultas eficientes. Este trabalho tem como objetivo realizar uma revisão sistemática da literatura para investigar as aplicações práticas de RAG e LLMs na construção de consultas SQL. A pesquisa busca identificar, organizar e analisar estudos existentes que explorem o uso desses modelos na criação e reestruturação de consultas, com o propósito de melhorar a eficiência de sistemas de banco de dados. A relevância do tema se destaca pela crescente necessidade de tornar a interação com bancos de dados mais acessível, permitindo que usuários sem conhecimento técnico possam realizar consultas complexas de forma intuitiva. Além disso, os avanços em inteligência artificial têm impulsionado a aplicação de modelos como RAG para otimizar processos de consulta e análise de dados. Por meio de uma revisão, busca-se mapear o estado da arte na área, destacando lacunas, avanços recentes e oportunidades para futuras pesquisas. A metodologia adotada segue o protocolo PRISMA, garantindo rigor na seleção, análise e síntese dos estudos incluídos. Foram utilizados critérios bem definidos para inclusão e exclusão de artigos, assegurando a confiabilidade dos achados. Os resultados indicam que as principais aplicações de RAG na construção de consultas SQL incluem assistentes de IA para consultas automatizadas, integração em chatbots corporativos, recuperação de informações em bases especializadas (como registros de saúde e propriedade intelectual) e otimização da geração de SQL em cenários complexos. Além disso, observam-se tendências promissoras, como o aprimoramento da precisão das consultas por meio da recuperação de contexto relevante e a redução do custo computacional para viabilizar o uso dessas técnicas em tempo real. Apesar do grande potencial, desafios técnicos ainda persistem, como a dificuldade na adaptação a cenários reais, a interpretação de consultas complexas e a necessidade de maior eficiência computacional.

Palavras-chave: Retrieval-Augmented Generation (RAG), Modelos de Linguagem de Grande Escala (LLMs), Consultas SQL, Bancos de Dados, Inteligência Artificial, Processamento de Linguagem Natural (PLN), Revisão Sistemática de Literatura, Eficiência em Sistemas de Informação.

ABSTRACT

Large Language Models (LLMs) are artificial intelligence models trained on vast amounts of textual data, capable of generating coherent responses and performing various natural language processing (NLP) tasks. The Retrieval-Augmented Generation (RAG) technique enhances these models by incorporating external knowledge sources, enabling them to generate more precise and up-to-date responses. In the context of databases, converting natural language into SQL queries (text-to-SQL) is a challenge, as it requires correctly interpreting the user's intent and generating efficient queries. This study aims to conduct a systematic literature review to investigate the practical applications of RAG and LLMs in constructing SQL queries. The research seeks to identify, organize, and analyze existing studies that explore the use of these models in creating and restructuring queries to improve database system efficiency. The relevance of this topic stands out due to the growing need to make database interaction more accessible, allowing users without technical knowledge to perform complex queries intuitively. Furthermore, advances in artificial intelligence have driven the application of models like RAG to optimize query processing and data analysis. Through a review, the study aims to map the state of the art in the field, highlighting gaps, recent advances, and opportunities for future research. The methodology follows the PRISMA protocol, ensuring rigor in the selection, analysis, and synthesis of the included studies. Well-defined criteria for article inclusion and exclusion were applied to ensure the reliability of the findings. The results indicate that the main applications of RAG in SQL query construction include AI assistants for automated queries, integration into corporate chatbots, information retrieval in specialized databases (such as health records and intellectual property data), and optimizing SQL generation in complex scenarios. Additionally, promising trends are observed, such as improving query accuracy through relevant context retrieval and reducing computational costs to enable real-time application of these techniques. Despite its great potential, technical challenges still persist, such as difficulty adapting to real-world scenarios, interpreting complex queries, and the need for greater computational efficiency.

Keywords: Retrieval-Augmented Generation, Large Language Models (LLMs), SQL Queries, Databases, Artificial Intelligence, Natural Language Processing (NLP), Systematic Literature Review, Information Systems Efficiency.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

UFPE	Universidade Federal de Pernambuco
LLM	Large Language Model
SQL	Structured Query Language
RSL	Revisão Sistemática de Literatura
NLP	Natural Language Processing
IA	Inteligência Artificial
AI	Artificial Inteligence
RAG	Retrieval-Augmented Generation
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SUMÁRIO

1 INTRODUÇÃO

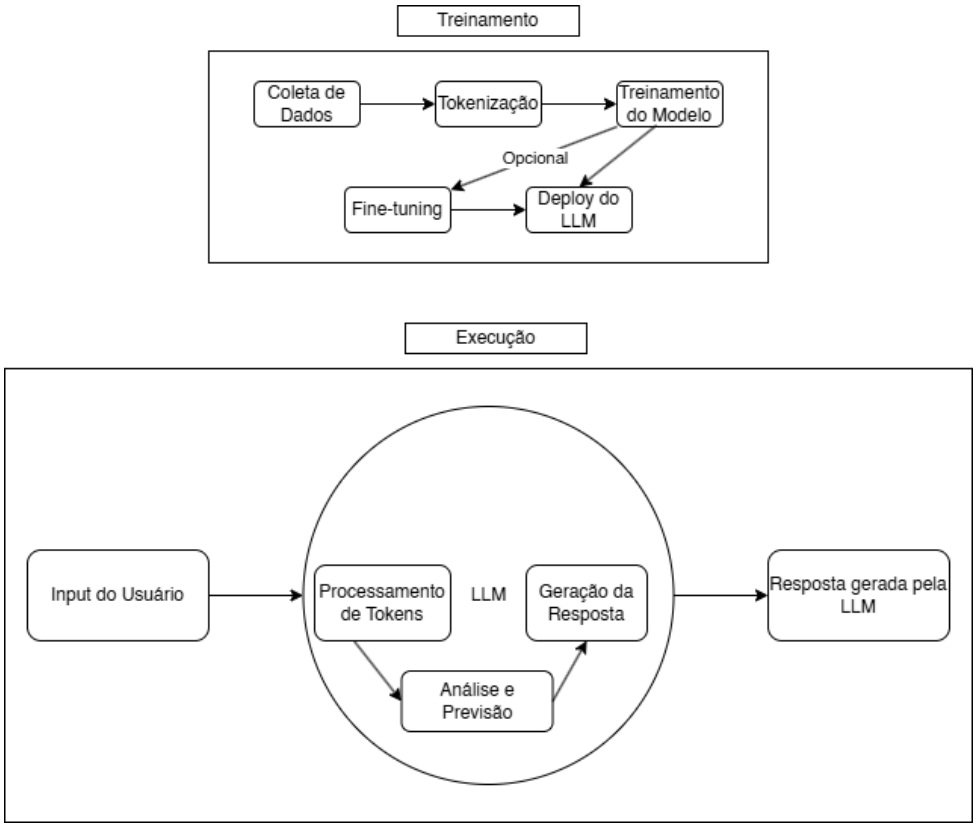
Os Modelos de Linguagem de Grande Escala, ou Large Language Models (LLMs), são modelos avançados de Inteligência Artificial (IA) treinados em quantidades vastas de dados em texto, usados para gerar outputs que simulam linguagem humana (ZHAO et al., 2023) [1]. Esses LLMs têm se destacado como ferramentas essenciais na conversão de linguagem natural em códigos, oferecendo resultados eficientes e precisos em diversas áreas (LI et al., 2024) [2]. Esses modelos, treinados em grandes corpora, demonstram capacidades notáveis em diversas tarefas de Processamento de Linguagem Natural (PLN) e exibem habilidades especiais que não estão presentes em modelos menores (ZHAO et al., 2023) [1].

Os Modelos de Linguagem de Grande Escala (LLMs) funcionam com base em arquiteturas de redes neurais profundas, especialmente o Transformer, que permite processar grandes volumes de dados textuais de forma eficiente. Esses modelos são pré-treinados nesses corpora massivos, utilizando técnicas como a modelagem de linguagem (language modeling), onde o objetivo é prever a próxima palavra em uma sequência de texto, capturando assim padrões linguísticos e conhecimentos gerais. Durante o pré-treinamento, os LLMs aprendem a gerar representações contextuais, que podem ser ajustadas posteriormente para tarefas específicas por meio de técnicas como fine-tuning e ajuste de instruções (instruction tuning). Além disso, os LLMs exibem habilidades emergentes, como aprendizado em contexto (in-context learning) e raciocínio passo a passo (step-by-step reasoning), que são ativadas quando o modelo atinge uma escala de parâmetros significativa (ZHAO et al., 2023) [1]. Essas capacidades permitem que os LLMs sejam aplicados em uma variedade de tarefas complexas, desde geração de texto até resolução de problemas matemáticos e programação, demonstrando uma versatilidade que os torna ferramentas poderosas no campo da IA.

A Retrieval-Augmented Generation (RAG), que pode ser traduzida como Geração Aumentada por Recuperação, é uma técnica que aprimora os modelos de linguagem de grande escala (LLMs) incorporando fontes de conhecimento externas, abordando limitações como informações desatualizadas e conteúdo impreciso (Lyu et al., 2024) [3]. O RAG geralmente envolve um processo em duas etapas: recuperar informações relevantes com base em uma consulta de entrada e gerar um texto informado tanto pela consulta

quanto pelo conhecimento recuperado (Shahade e Deshmukh, 2024) [4].

Figura 1: Funcionamento de Large Language Models.

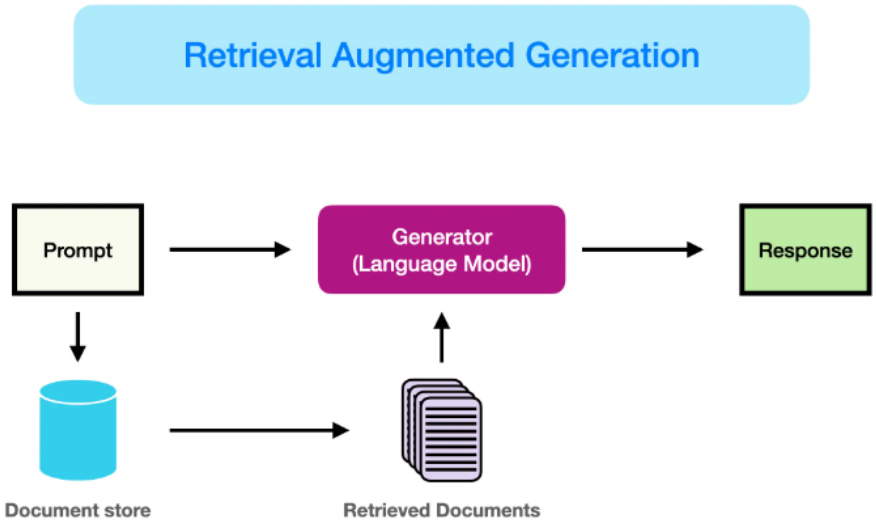


Fonte: De autoria própria.

A Figura 1 destaca as duas principais fases do funcionamento de um modelo de Large Language Model (LLM): treinamento e execução. Na primeira parte do diagrama, referente ao treinamento, observa-se o fluxo que envolve a coleta de dados, onde o modelo é alimentado com textos extraídos de diversas fontes, seguida da tokenização, que transforma esses textos em unidades menores processáveis pelo modelo. A etapa seguinte, o treinamento do modelo, utiliza redes neurais para ajustar os pesos e aprender padrões a partir dos dados processados. Caso necessário, realiza-se o fine-tuning, um refinamento do modelo em domínios específicos para melhorar sua precisão. Após essa fase, o modelo passa pelo deploy, onde a versão treinada é disponibilizada para uso real. Já a segunda parte da figura ilustra a fase de execução, que ocorre quando um usuário interage com o modelo. O input do usuário, geralmente uma pergunta ou comando em linguagem natural, é recebido pelo LLM e passa pelo processamento de tokens, onde a entrada é segmentada para análise. Em seguida, ocorre a análise e previsão, na qual o modelo processa o contexto da pergunta e gera uma resposta provável. Esse resultado passa pela fase de

geração de resposta, refinando a saída antes de apresentá-la ao usuário. Finalmente, a resposta gerada pela LLM é entregue como saída.

Figura 2: Funcionamento de Retrieval-Augmented Generation.



Fonte: Prompt Engineering Guide, 2025. [5].

A Figura 2 ilustra o funcionamento da técnica de Retrieval-Augmented Generation (RAG) e sua interação com um Modelo de Linguagem de Grande Escala (LLM). No diagrama, observa-se que o RAG atua como uma camada de busca especializada, responsável por recuperar informações relevantes a partir de uma base de conhecimento externa. Esse processo é essencial para fornecer ao LLM um contexto mais atualizado e preciso, reduzindo a dependência exclusiva do conhecimento pré-treinado do modelo (Lyu et al., 2024) [3]. O fluxo representado na imagem inicia-se com uma consulta em linguagem natural, que é processada pelo RAG para identificar e recuperar dados complementares. Esses dados são então incorporados como contexto adicional antes que o LLM gere sua resposta final. Com essa abordagem, o modelo pode fundamentar suas respostas em informações mais específicas e recentes, mitigando problemas como alucinações e limitações de conhecimento estático.

Dentre as aplicações de LLMs, destaca-se o text-to-SQL, uma tarefa que converte input textual em consultas SQL (Structured Query Language), com objetivo de usar a linguagem natural para interação com sistemas de bancos de dados (DENG et al., 2022) [6]. Essa abordagem de text-to-SQL enfrenta desafios importantes, como a codificação

da semântica da linguagem do usuário, a geração precisa de comandos SQL e a conexão entre a intenção do usuário e sua implementação prática (GUO et al., 2019) [7]. Essas dificuldades tornam o uso de LLMs nesse contexto um campo de grande potencial e complexidade, reforçando a importância de estudos que explorem suas capacidades e limitações.

Segundo Jeon et al. (2023) [8], com o avanço de modelos como o GPT-4 e sua popularização, os LLMs têm ganhado relevância, destacando a necessidade de aprofundar o estudo sobre suas aplicações. Em particular, a SQL, por ser uma linguagem declarativa — em contraste com as linguagens imperativas tradicionais —, potencializa o uso de LLMs, permitindo a geração de consultas mais eficazes e alinhadas às necessidades dos usuários (YE et al., 2023) [9]. Ademais, já existem aplicações sendo desenvolvidas com o objetivo de usar LLMs como interface para bases de dados, evidenciando o avanço recente na área e indicando potenciais futuras implementações.(LI et al., 2024) [2].

Pesquisas recentes têm se concentrado no aprimoramento da Geração Aumentada por Recuperação para o processamento de consultas estruturadas e a geração de SQL. Yang et al. (2024) [10] introduziram a Integração Semântica Generativa para melhorar o desempenho do RAG na compreensão de consultas estruturadas. Ziletti e D'Ambrosi (2024) [11] desenvolveram uma abordagem baseada em RAG para a geração de text-to-SQL em respostas a perguntas epidemiológicas usando registros eletrônicos de saúde. Yang et al. (2024) [10] propuseram o RAG Baseado em Consultas (QB-RAG), que pré-computa consultas potenciais para melhorar a precisão das respostas em questões de saúde. Allu et al. (2024) [12] abordaram consultas complexas em tabelas de PDFs, extraindo e enriquecendo o conteúdo tabular antes da sumarização. Esses estudos demonstram o potencial do RAG em diversos domínios, incluindo saúde e análise de dados. Embora os modelos de linguagem atuais possam não ser suficientemente precisos para uso não supervisionado, o RAG oferece perspectivas promissoras para aprimorar suas capacidades no processamento de consultas estruturadas e na geração de SQL (Ziletti e D'Ambrosi, 2024 [11]; Yang et al., 2024 [10]).

A motivação para este estudo reside nesse crescente impacto dos modelos de linguagem de grande escala em aplicações práticas, especialmente no contexto de bancos de dados. Com a explosão de dados e o crescimento da demanda por ferramentas que facilitem a interação entre usuários e sistemas, os LLMs têm se mostrado promissores

na tarefa de text-to-SQL, permitindo que pessoas sem conhecimentos técnicos avancem consultas complexas apenas utilizando linguagem natural, além de servir como otimizador para atuantes mais experientes na área. Além disso, a implementação de técnicas como o RAG mostraram-se capazes de potencializar o uso dessas LLMs, tornando-as mais confiáveis, expandindo a capacidade de criação de resultados ainda mais próximos aos desejados pelo usuário. Por exemplo, a inclusão de indicadores de citação no texto proporciona diversos benefícios claros, como a possibilidade desses usuários verificarem facilmente as afirmações feitas pelos LLMs com base nas referências fornecidas, melhorando assim a transparência e a credibilidade do conteúdo. (Lyu et al., 2024) [3]. Além disso, nos casos de aplicação de RAG como potencializador de LLMs em tarefas de text-to-SQL, há um grande ganho em adaptabilidade do modelo para o banco, especialmente em cenários onde a estrutura dos dados é altamente diversa e menos padronizada. Diferente dos benchmarks acadêmicos, que costumam conter esquemas bem definidos e simplificados, bancos de dados reais apresentam variações complexas, como múltiplas tabelas inter-relacionadas, esquemas dinâmicos e metadados incompletos, tornado a aplicação de RAG um diferencial na implementação. [13]

Apesar do grande potencial, desafios técnicos ainda dificultam a implementação desses modelos em larga escala. Entre eles, destacam-se a geração precisa de consultas SQL, a escalabilidade para cenários reais e a interpretação de intenções ambíguas, ou seja, solicitações que podem ter múltiplas interpretações dependendo do contexto. Ademais, a ausência de padrões bem definidos e a escassez de estudos consolidados limitam a compreensão sobre como otimizar o desempenho desses modelos e mitigar suas limitações [3]. Portanto, compreender os desafios técnicos, como o aumento do custo computacional e a dificuldade na interpretação de consultas complexas, e as tendências emergentes, como otimização do contexto recuperado para reduzir alucinações e o aprimoramento da adaptação dos modelos a cenários reais, dentro do uso de RAG e LLMs no text-to-SQL, não apenas preenche uma lacuna crítica no estado da arte, mas também contribui diretamente para avanços na acessibilidade, eficiência e inovação em sistemas de banco de dados. Então, com o objetivo de investigar as aplicações e os desafios técnicos dessa abordagem, este estudo visa aprofundar o entendimento das capacidades e limitações dos RAG em contextos reais. Para este fim, será realizada uma Revisão Sistemática da literatura sobre a aplicação de RAG com SQL, utilizando o protocolo Principais Itens para Relatar Revisões

Sistemáticas e Meta-análises (PRISMA)(MOHER et al., 2009) [14].

2 METODOLOGIA

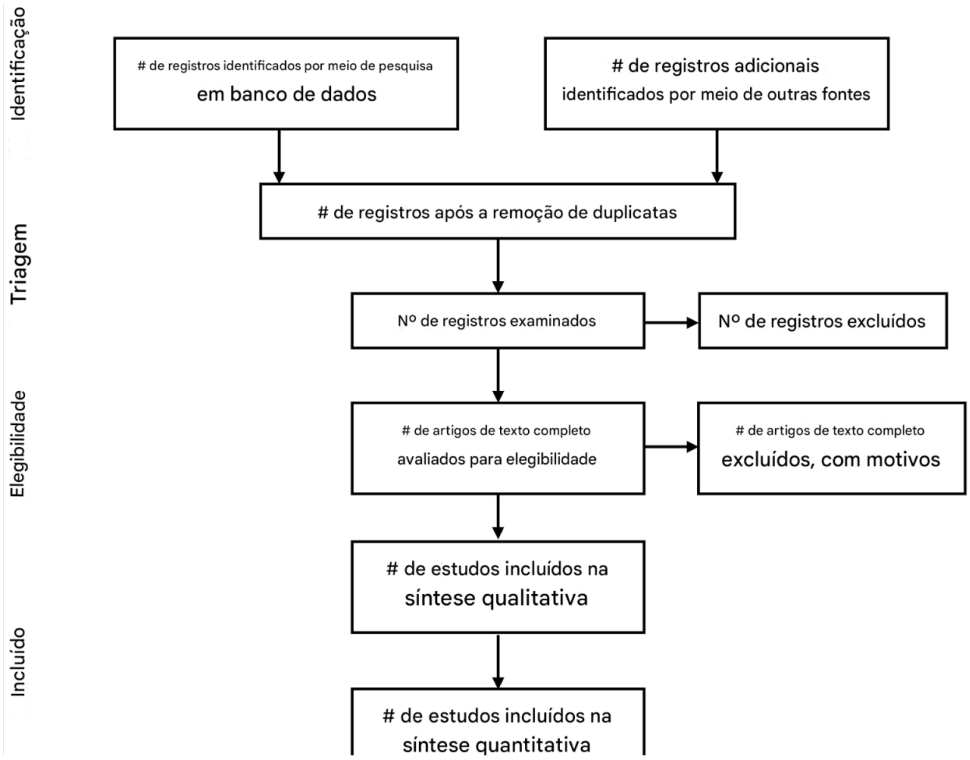
2.1 PRISMA

A revisão sistemática deste estudo será conduzida de acordo com o protocolo PRISMA (MOHER et al., 2009) [14], que visa garantir a transparência e rigorosidade em todas as etapas da pesquisa. O PRISMA proporciona uma abordagem padronizada para a seleção, análise e síntese dos estudos incluídos, assegurando que cada fase do processo seja realizada de maneira clara e replicável. Ao seguir esse protocolo, busca-se minimizar o viés na seleção de estudos e promover a consistência na avaliação da qualidade metodológica das pesquisas envolvidas. O uso do PRISMA também facilita a compreensão e a reprodutibilidade dos resultados, promovendo uma maior confiança nas conclusões da revisão sistemática. O método assegura que a pesquisa seja conduzida com altos padrões científicos, levando em consideração a inclusão de critérios bem definidos para a escolha dos estudos, a avaliação da qualidade e a extração de dados relevantes, o que é fundamental para garantir a robustez e a validade dos achados da revisão. Abaixo, na Figura 3, o fluxograma das fases de uma revisão sistemática da literatura.

2.1.1 **Definição do Problema de Pesquisa**

Em primeira análise, foi necessário definir os problemas de pesquisa. Para isso, foram formuladas de maneira clara e objetiva, com o intuito de direcionar todas as etapas subsequentes da revisão. As perguntas orientam a busca por estudos, os critérios de inclusão e exclusão e a análise dos resultados. Para garantir a objetividade da pesquisa, a questão foi estruturada utilizando o modelo PICO (População, Intervenção, Comparação e Desfecho), adaptado para o contexto do estudo de Modelos de Linguagem de Grande Escala (LLMs) aplicados à construção de consultas SQL (consulte a Tabela 1).

Figura 3: Fluxograma das fases de uma RSL



Fonte: Traduzido de Moher (2009, p. 3) [14].

Tabela 1: Modelo PICO

Acrônimo	**Definição**	**Descrição**
P	População ou problema	Estudos que envolvam a aplicação de RAG (Retrieval-Augmented Generation) e LLMs em sistemas de banco de dados, com foco em construção e reestruturação de consultas SQL.
I	Intervenção	Uso de RAG para melhorar a eficiência e precisão na criação e modificação de consultas SQL.
C	Comparação	Métodos tradicionais ou outras abordagens computacionais para a construção de consultas SQL, como ferramentas específicas de bancos de dados ou algoritmos convencionais de NLP.
O	Resultado	Identificação de melhorias na eficiência, precisão e aplicabilidade no uso de RAG e LLMs em consultas SQL, além do mapeamento do estado da arte e de lacunas na literatura.

O objetivo é identificar as principais aplicações, desafios técnicos, ferramentas utilizadas e comparar o desempenho dos RAG e LLMs com abordagens tradicionais, tendo cada problema de pesquisa uma motivação elencada (consulte a Tabela 2).

Tabela 2: Perguntas de Pesquisa (Research Questions ou RQ) e Motivações

Pergunta de Pesquisa	**Motivação**
RQ1. Que tipo de contribuição e objetivo de pesquisa estão sendo abordados nos estudos?	Entender as contribuições de cada pesquisa permite não só mapear as áreas de aplicação, mas também as limitações enfrentadas pelos modelos e as possíveis melhorias propostas.
RQ2. Quais são as principais aplicações de RAG na construção de consultas SQL?	Identificar como RAG e LLMs estão sendo aplicados na prática pode revelar oportunidades de inovação, exemplos de sucesso e áreas de impacto nos sistemas de banco de dados.
RQ3. Quais desafios técnicos têm sido relatados no uso de RAG na construção de consultas SQL?	Compreender os obstáculos técnicos é crucial para orientar melhorias futuras, como otimização de modelos, redução de erros sintáticos/semânticos e aumento da compatibilidade com diferentes sistemas.
RQ4. Como o uso de RAG potencializa os LLMs em tarefas de text-to-SQL?	Investigar como RAG aprimora LLMs em text-to-SQL é essencial para avaliar seu impacto na precisão das consultas, na adaptação a diferentes bancos de dados e na superação das limitações dos modelos tradicionais.

2.1.2 **Planejamento e Delimitação de Escopo**

Os objetivos da revisão estão diretamente relacionados às questões de pesquisa e visam aprofundar o entendimento sobre as capacidades e limitações dos LLMs no contexto do text-to-SQL e como o RAG é implementado. Buscou-se identificar as aplicações mais frequentes desses modelos, os desafios técnicos mais comuns enfrentados pelos pesquisadores, e avaliar ferramentas e frameworks utilizados. O escopo da revisão foi delimitado para incluir estudos que abordem o uso de RAG e LLMs especificamente para a tarefa de text-to-SQL, ou seja, a conversão de linguagem natural em comandos SQL. A revisão foca em estudos que investigam a aplicação desses modelos em bancos de dados, considerando aspectos como precisão, eficiência e escalabilidade. Estudos fora desse escopo, ou que envolvem outros tipos de modelos de IA ou outras aplicações de SQL, foram excluídos.

2.1.3 **Estratégia de Busca**

Somado a isso, então, foi feita a escolha de bases de dados e fontes que melhor atendem os objetivos de pesquisa, sendo utilizado o Google Scholar, reconhecido por sua ampla acessibilidade, abrangência e variedade de artigos acadêmicos relevantes, o que garante uma análise consistente e alinhada com o estado da arte na área. Foi feito também o levantamento de uma string de busca que melhor abrangesse estudos potencialmente relevantes e restringisse artigos que seriam descartados, utilizando sinônimos com o operativo OR e artigos interseccionais com o operativo AND, chegando na string:
("Large Language Models"OR "LLMs"OR "LLM"OR "Large-Scale Language Models") AND ("SQL"OR "Structured Query Language") AND ("text-to-SQL"OR "natural language to SQL") AND ("RAG" OR "Retrieval Augmented Generation") AND ("optimization"OR "performance improvement")

2.1.4 **Critérios de Inclusão e Exclusão**

Os critérios de inclusão foram definidos para garantir que os estudos selecionados abordem o uso de RAG e LLMs na construção de consultas SQL, com foco em estudos que ofereçam dados sobre a eficácia, aplicabilidade e limitações desses modelos. Foram incluídos, por exemplo, apenas estudos publicados em periódicos e conferências de alta qualidade e relevância. Os critérios de exclusão envolveram a remoção de estudos que não apresentassem dados quantitativos ou qualitativos sobre a aplicação de RAG no contexto de text-to-SQL, além de artigos que não tratassem diretamente das questões relacionadas à construção de consultas SQL (consulte a Tabela 3).

Tabela 3: Critérios de Inclusão e Exclusão

#	**Critérios de Inclusão**
1	Estudos que abordem RAGs aplicados à construção de consultas SQL.
2	Estudos que tratem de avanços técnicos, desafios ou ferramentas utilizadas no uso de RAGs em SQL.
3	Estudos científicos, revisões, ou capítulos de livro publicados em revistas acadêmicas ou conferências reconhecidas.
4	Estudos publicados entre 2021 e 2025.
5	Trabalhos publicados em inglês.
6	Estudos com texto completo disponível para análise (PDFs acessíveis).
#	**Critérios de Exclusão**
1	Artigos que não mencionam RAGs ou SQL diretamente.
2	Estudos focados em consultas SQL, mas sem conexão com aprendizado de máquina ou RAGs.
3	Artigos sem revisão por pares.
4	Literatura Cinzenta.
5	Short Papers (menores que 5 páginas).
6	Trabalhos opinativos ou sem evidências experimentais ou metodológicas claras.
7	Artigos publicados de forma anônima.
8	Artigos duplicados.

Os critérios de idioma adotados na revisão foram definidos para assegurar o pleno entendimento dos artigos selecionados, evitando potenciais limitações decorrentes de traduções que poderiam comprometer a acurácia na análise e interpretação dos dados. Já o período buscou avaliar tanto o crescimento recente de pesquisas na área ao mesmo tempo que evitar análises de RAG e LLMs mais antigas que ainda não tivessem a eficácia de modelos atuais, tornando o estudo mais defasado e heterogêneo.

2.1.5 **Processo de Seleção de Estudos**

Durante o processo de planejamento, foram identificados alguns desafios potenciais, como a heterogeneidade dos estudos encontrados, a qualidade variável das metodologias e a limitação de estudos sobre a comparação direta entre RAG e abordagens tradicionais de construção de SQL. Também foi identificado com risco potencial a utilização de artigos com objetivo maior de divulgação de um produto. Para mitigar esses riscos, foi estabelecido um processo rigoroso de seleção dos estudos e análise crítica de sua metodologia, garantindo que apenas os estudos mais relevantes e de maior qualidade fossem incluídos na revisão.

2.1.6 **Extração e Síntese de Dados**

Após a seleção dos estudos incluídos na revisão sistemática, foi realizada a etapa de extração e síntese dos dados. Esse processo envolveu a leitura detalhada dos artigos selecionados, com o objetivo de identificar e organizar as informações mais relevantes para responder às perguntas de pesquisa. Durante a extração, foram registrados aspectos como a metodologia utilizada nos estudos, os principais desafios técnicos relatados, as abordagens de implementação do RAG em consultas SQL e as tendências emergentes na área.

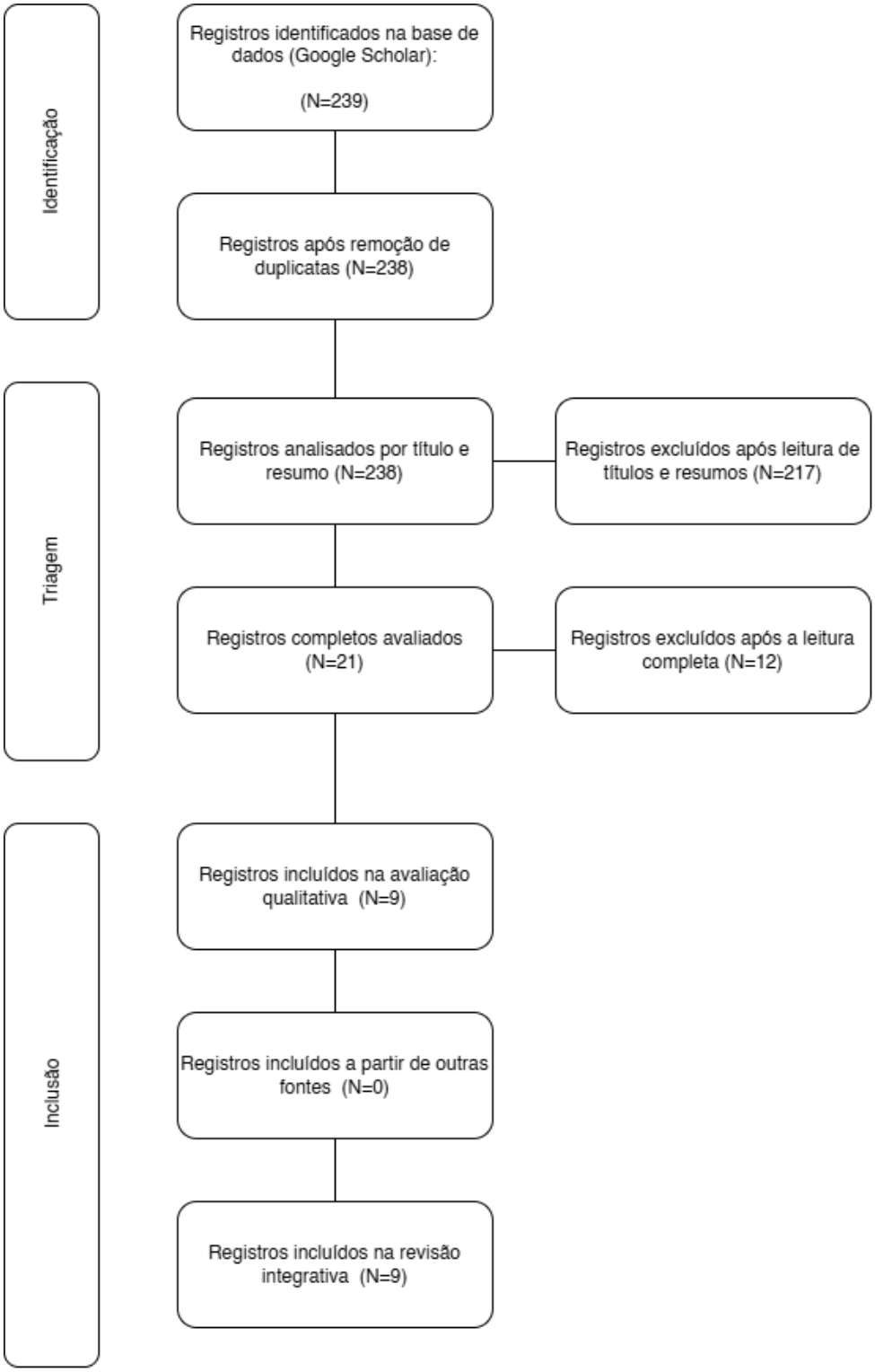
A síntese dos dados ocorreu por meio da categorização das informações coletadas, agrupando os estudos de acordo com seus principais focos e contribuições. Esse procedimento permitiu identificar padrões, lacunas e oportunidades de aprimoramento na utilização de LLMs e RAG na geração de SQL. Além disso, aspectos como a precisão das consultas geradas, a influência de diferentes bases de conhecimento e os impactos da técnica na eficiência de bancos de dados foram documentados, garantindo uma análise estruturada e aprofundada sobre o tema.

3 RESULTADOS

Na etapa de triagem dos artigos, foi realizada uma análise com base nos critérios de exclusão estabelecidos. Dos 239 artigos inicialmente identificados, a maioria foi excluída por não atender aos requisitos da pesquisa. Especificamente, 113 artigos foram provenientes do repositório arXiv, que não conta com revisão por pares, e foram, portanto, descartados. Sete artigos não eram efetivamente artigos científicos, e 69 tangenciam apenas parcialmente os tópicos de interesse, sendo excluídos por não abordarem diretamente o assunto. Além disso, outros 7 artigos estavam em idiomas diferentes do inglês, e 13 apresentavam problemas de acesso, como PDFs indisponíveis. Oito artigos eram voltados para divulgação de produtos ou frameworks e não contribuiriam com a natureza acadêmica do trabalho, enquanto um artigo foi repetido. Após aplicar esses critérios, 21 artigos foram selecionados para leitura completa, representando as fontes mais relevantes e adequadas aos objetivos da pesquisa, e se mostraram aptos para etapa de avaliação da elegibilidade.

Após a leitura completa dos 21 artigos selecionados, foram aplicados os critérios para determinar quais estudos seriam considerados na revisão sistemática. Como resultado, nove artigos foram escolhidos para seguir na pesquisa, garantindo que apenas os trabalhos mais relevantes e alinhados aos objetivos do estudo fossem incluídos na análise final, com estudos sendo descartados por menor enfoque nos paradigmas de aplicação de RAG para construção de consultas SQL, ou tangenciando o tema ou se mostrando uma pesquisa científica mais voltada para divulgação de produtos e frameworks. A Figura 4 apresenta esse processo de seleção, mostrando o fluxo de escolha de artigos, enquanto a Tabela 4 apresenta as informações dos artigos escolhidos, contendo os dados dos artigos: ordem, ano, autores e título.

Figura 4: Fluxo do processo de seleção dos artigos.



Fonte: elaborado de acordo com o modelo de Moher et al. (2009) [14].

Tabela 4: Síntese dos Estudos Analisados

N°	**Ano**	**Autores**	**Título**
1	2024	Syrjä, Saku-Matti	Retrieval-Augmented Generation Utilizing SQL Database – Case: Web Sport Statistics Application
2	2024	Orrenius, Axel	Enhancing Text-to-SQL Applications with Retrieval Augmented Generation
3	2024	Bartczak, Zuzanna	From RAG to Riches: Evaluating the Benefits of Retrieval-Augmented Generation in SQL Database Querying
4	2024	Henriques, Pedro Duarte Santos	Augmenting Large Language Models with Context Retrieval
5	2024	Rossi, Alessandro	Large Language Models to Query Your Data: Retrieving Ads Industry Users Data Using Natural Language
6	2024	Rizzi, Fabio	Developing an Enterprise Chatbot using Machine Learning Models: A RAG and NLP-based Approach
7	2024	Maj, Michał; Pliszczuk, Damian; Marek, Patryk; Wilczewska, Weronika; Przysucha, Bartosz; Rymarczyk, Tomasz	Optimizing Customer Support Using Text2SQL to Query Natural Language Databases
8	2024	Chiras, Marios	Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy
9	2024	Eriksson, Alice	Chat-based Search for Intellectual Property Data

A totalidade dos artigos selecionados ser de 2024 evidencia que o tema ainda é recente e está em fase de desenvolvimento, com tendência de crescimento nos próximos anos. Além disso, a quantidade de artigos excluídos por tratarem de frameworks ou produtos comerciais ressalta a necessidade de um aprofundamento acadêmico na área, dissociado de interesses empresariais. Outro ponto relevante é que grande parte dos artigos descartados por falta de revisão por pares também eram de 2024, o que indica que a temática ainda não teve ampla disseminação em conferências e periódicos científicos. No entanto, essa limitação tende a ser reduzida à medida que o tema amadurece e passa a ser abordado em eventos acadêmicos e pesquisas futuras na área de tecnologia.

3.1 QUE TIPO DE CONTRIBUIÇÃO E OBJETIVO DE PESQUISA ESTÃO SENDO ABORDADOS NOS ESTUDOS?

O objetivo desta pergunta é entender as diferentes contribuições realizadas por cada autor nessa revisão sistemática de literatura, sumarizando o que foi encontrado em cada pesquisa e buscando direcionar onde estarão as maiores contribuições de cada trabalho.

O estudo de Syrjä [15] investiga a aplicação de Retrieval-Augmented Generation (RAG) para automatizar a criação de consultas SQL a partir de linguagem natural, tornando os bancos de dados mais acessíveis a usuários não técnicos. A pesquisa demonstra essa abordagem por meio do desenvolvimento de um sistema de estatísticas esportivas que utiliza um modelo de linguagem (LLM) para interpretar perguntas em linguagem natural e convertê-las em consultas SQL precisas. Por exemplo, ao receber uma pergunta como "Quantos gols McDavid marcou na temporada 2023-2024?", o sistema gera automaticamente a consulta correspondente no banco de dados. Esse processo elimina a necessidade de conhecimento prévio em SQL, permitindo que qualquer usuário acesse informações estruturadas de forma intuitiva. No entanto, o estudo também identifica desafios técnicos, como a dificuldade dos modelos em lidar com a formatação de temporadas esportivas e a seleção correta de colunas na base de dados, evidenciando limitações na integração entre LLMs e bancos de dados relacionais.

Rossi [16] expande a aplicação de modelos RAG para a automação da geração de consultas SQL no ambiente corporativo, focando na otimização de processos internos em empresas que lidam com grandes volumes de dados e consultas complexas. Sua pesquisa apresenta um caso de uso na indústria de publicidade digital, onde a plata-

forma Advertising Resource Management (ARM) enfrenta desafios relacionados à alocação de orçamentos publicitários em múltiplos canais, exigindo consultas SQL precisas para análise de desempenho. O estudo propõe a integração de um assistente de IA baseado em LLMs, permitindo que usuários solicitem dados em linguagem natural, como "Quais campanhas tiveram maior ROI no último trimestre?", e obtenham respostas diretas sem necessidade de conhecimento técnico. Além disso, Rossi destaca o uso do algoritmo Max Marginal Relevance (MMR) para recuperar exemplos relevantes e melhorar a geração de SQL, além da implementação de um modelo de segurança para evitar consultas maliciosas. Os resultados indicam ganhos em precisão e engajamento dos usuários, reforçando o potencial dos LLMs na automação de processos empresariais.

A pesquisa de Rizzi [17] foca na aplicação de RAG para a geração de consultas SQL em tempo real a partir de linguagem natural, facilitando a interação com bases de dados empresariais. No contexto do artigo, essa abordagem foi implementada em um chatbot corporativo para a empresa Betacom S.R.L., onde funcionários sem conhecimento técnico podem obter informações de bancos de dados sem precisar formular consultas SQL manualmente. Por exemplo, um gerente de compras que precise de informações sobre transações armazenadas no banco de dados pode simplesmente perguntar ao chatbot em linguagem natural. O sistema então utiliza um modelo baseado em RAG para recuperar os dados relevantes e, com o auxílio do modelo PipableAI, gerar uma consulta SQL precisa. Essa consulta é então executada e os resultados são apresentados ao usuário de forma tabular ou gráfica, dependendo da necessidade. Outrossim, o chatbot também permite a busca em documentos internos da empresa, utilizando um modelo LLama 3 8B quantizado, FAISS para recuperação semântica e BM25 para buscas sintáticas. Dessa forma, a solução de Rizzi melhora a produtividade empresarial, reduzindo a dependência de especialistas em SQL e otimizando o acesso às informações essenciais para a tomada de decisões.

Já Chiras [18] propõe uma aplicação especializada de RAG no campo da fisioterapia esportiva, mais especificamente em análises de dados biomecânicos. Por exemplo, um fisioterapeuta que acompanha a reabilitação de um atleta pode simplesmente perguntar ao sistema se o desempenho muscular do paciente está dentro dos padrões esperados. O agente, então, usa um modelo de RAG para recuperar dados biomecânicos relevantes e, com o auxílio de um modelo de Text-To-SQL, gera e executa uma consulta para extrair as métricas correspondentes. Os resultados são analisados e apresentados ao profissional

de forma clara, destacando possíveis anomalias, como esforço excessivo em determinados grupos musculares. A pesquisa de Chiras também avaliou diferentes tamanhos de modelos de linguagem, analisando o equilíbrio entre precisão e consumo de recursos computacionais. Foi constatado que modelos maiores, como Llama3:70b, oferecem maior precisão na detecção de outliers nos dados biomecânicos, mas demandam mais poder computacional. A solução proposta demonstra a flexibilidade dos modelos RAG na adaptação a domínios técnicos complexos, permitindo mais precisão na análise fisioterapêutica e facilitando a tomada de decisões clínicas.

Em sua pesquisa, Eriksson [19] apresenta uma aplicação também especializada, utilizando RAG para melhorar a recuperação de informações em bancos de patentes. Seu estudo mostra como esses modelos podem ser empregados no contexto da propriedade intelectual, visando aprimorar o processo de pesquisa e consulta de dados relacionados a patentes. No contexto do artigo, o sistema desenvolvido por Eriksson foi implementado no Swedish Intellectual Property Office (PRV) e utiliza um chatbot baseado em RAG para permitir consultas interativas e mais eficientes. Um pesquisador, por exexmplo, que precise verificar se uma nova invenção já foi patenteada pode simplesmente perguntar ao chatbot em linguagem natural. O sistema, então, usa um modelo de recuperação híbrido, combinando busca semântica e NL2SQL, um tipo especializado de text-to-SQL, para recuperar informações relevantes do banco de dados de patentes. Em seguida, um modelo GPT-4o gera uma resposta contextualizada, garantindo precisão e relevância. Esse processo evita a necessidade de os usuários formularem consultas complexas ou navegarem por múltiplos documentos técnicos. O estudo também avaliou a eficiência do chatbot por meio de testes com usuários, indicando uma melhoria na precisão das buscas e na experiência de consulta em comparação com sistemas tradicionais baseados em palavras-chave. A contribuição de Eriksson está em explorar o uso de RAG em áreas com grandes volumes de dados, exigindo alta especialização e precisão nas consultas, além de demonstrar o potencial de interfaces conversacionais para simplificar o acesso à informação em bases de propriedade intelectual.

Orrenius [13] realiza uma avaliação do impacto dos modelos RAG na precisão das consultas SQL, particularmente em ambientes mais complexos. Sua pesquisa destaca o potencial desses modelos para melhorar a acurácia das consultas com base em técnicas como in-context learning e engenharia de prompts. O estudo utiliza benchmarks como

Spider e um dataset real da empresa Nibiru Software para medir a eficácia. Para aumentar a precisão, Orrenius aplica estratégias de calibração de viés, inserindo dicas e exemplos que orientam o LLM a evitar consultas excessivamente amplas e a limitar o uso de operadores complexos, como LEFT JOIN e OR. Além disso, utiliza votação majoritária para validar a consulta mais adequada dentre múltiplas geradas. A contribuição de Orrenius está em fornecer uma análise crítica que embora os modelos demonstrem resultados promissores em benchmarks, há desafios em transferir esses avanços para contextos reais, como discrepâncias na estrutura de dados e na semântica das perguntas. Orrenius também aborda a necessidade de ajustes finos, que incluem a otimização de prompts, a integração de correção automática de SQL gerados com erros e o uso de vetores embutidos para facilitar a seleção de tabelas e colunas relevantes. Essas intervenções mitigam problemas típicos, como alucinações do modelo e consultas malformadas, melhorando a eficácia geral das consultas SQL em bases de dados não padronizadas.

Já Bartczak [20] avalia o uso do RAG em diferentes LLMs para consultas SQL, conduzindo testes empíricos com modelos como GPT-3.5, GPT-4, Llama2 e Llama3. Os resultados indicam que, embora o uso de RAG melhore a precisão das consultas, isso é, resultados com consultas SQL que satisfaçam a necessidade do usuário, ele também aumenta o custo computacional, chegando a aumentar em até 250% em consultas mais complexas, o que limita a sua aplicação em tempo real. Além disso, a pesquisa propõe uma abordagem mais eficiente para a implementação de RAG, destacando-se a redução do tamanho do contexto recuperado, filtrando apenas informações essenciais para a geração da consulta SQL. Além disso, a pré-indexação de esquemas de banco de dados com embeddings permite otimizar a recuperação de metadados, minimizando buscas redundantes..

Henriques [21] propõe um método de seleção de contexto em duas fases para otimizar a geração de consultas SQL usando LLMs. Sua pesquisa demonstra que a técnica reduz os custos computacionais ao empregar um modelo compacto na primeira fase para selecionar apenas as informações mais relevantes, minimizando o número de tokens processados pelo modelo principal na segunda etapa. Embora o ganho em acurácia tenha sido de apenas 2% na precisão das consultas SQL, a abordagem permitiu processar 30% mais instâncias, evidenciando sua eficiência na escalabilidade do sistema. Henriques destaca a importância dessa otimização para aplicações que exigem desempenho em tempo real, ressaltando que a redução de custos computacionais pode compensar a melhoria limitada

na precisão, especialmente em tarefas de geração de SQL.

Por fim, Maj et al. [22] investigam o impacto do excesso de contexto na precisão das consultas em sistemas de Text2SQL (ou text-to-SQL), utilizando Llama3, Gemma2 e Codegemma. Um exemplo de excesso de contexto ocorre quando o modelo recebe informações detalhadas demais sobre o esquema do banco de dados, incluindo descrições extensivas das tabelas e colunas. Nos testes realizados, a adição de informações mais detalhadas não melhorou a taxa de acerto dos modelos e, em alguns casos, reduziu sua acurácia e aumentou o tempo de resposta, caracterizando um efeito semelhante ao overfitting, que ocorre quando um modelo aprende padrões excessivamente específicos dos dados de treino, perdendo a capacidade de generalizar para novos dados. Para lidar com esse problema, a pesquisa propõe a redução do contexto, limitando as informações fornecidas ao modelo apenas ao essencial para a formulação da consulta SQL. Isso é feito por meio da seleção automática de metadados relevantes, como tabelas e colunas mais consultadas, reduzindo a carga de processamento sem comprometer a qualidade da resposta. Assim, a abordagem se concentra em fornecer um conjunto de informações mais seletivo e direcionado, garantindo que o modelo não seja sobrecarregado por detalhes desnecessários.

3.2 QUAIS SÃO AS PRINCIPAIS APLICAÇÕES PRÁTICAS DE MODELOS DE RAG NA CONSTRUÇÃO DE CONSULTAS SQL?

A pergunta busca entender onde os modelos de RAG têm sido utilizados na prática, tentando buscar reconhecer tendências. Os estudos analisados demonstram contribuições práticas do uso de Modelos de Geração Aumentada por Recuperação na construção de consultas SQL, com foco em diversas áreas e contextos de aplicação.

Modelos de RAG têm sido utilizados na criação automatizada de consultas SQL, especialmente em assistentes de IA, como os chatbots. Syrjä [15], Rossi [16] e Rizzi [17] exploram a aplicação de RAG para gerar consultas SQL a partir de linguagem natural, o que facilita a interação de usuários sem conhecimento técnico com bancos de dados, aumentando acessibilidade e rompendo barreiras técnicas.

A aplicação de RAG em contextos especializados também é uma área promissora. Chiras [18] utiliza RAG para gerar consultas SQL em análises de dados biomecânicos no campo da fisioterapia esportiva, enquanto Eriksson [19] aplica o modelo em dados de propriedade intelectual, especificamente para a recuperação de informações em ban-

cos de patentes. Essas abordagens mostram como RAG pode ser eficaz em setores que exigem consultas complexas e específicas. Ainda segundo Eriksson [19], é analisado que diferentemente de um LLM genérico, que pode gerar respostas imprecisas devido à falta de contexto especializado, o uso de RAG permite recuperar informações mais relevantes e direcionadas, aumentando a precisão das consultas em domínios altamente técnicos.

Ademais, RAG tem sido integrado em sistemas corporativos, especialmente em chatbots para consultas SQL. A pesquisa de Rizzi [17] demonstra como RAG pode ser utilizado para otimizar a interação com bases de dados empresariais por meio de chatbots. Implementado na empresa Betacom S.R.L., o modelo permite que funcionários consultem informações sem necessidade de conhecimento técnico em SQL, automatizando a formulação de consultas e melhorando a eficiência no acesso a dados estratégicos. Além disso, a integração com técnicas como recuperação semântica e buscas sintáticas amplia a aplicabilidade do sistema, tornando a consulta a informações corporativas mais ágil e acessível.

Orrenius [13] avalia a utilização de RAG para melhorar a geração de SQL em sistemas de consulta, com base em benchmarks como Spider e testes em dados reais da Nibiru Software. Sua pesquisa confirma que RAG aprimora a precisão das consultas SQL, especialmente em interações mais complexas com bancos de dados. No entanto, ele destaca que a transição dos modelos para ambientes reais exige um esforço adicional, pois diferenças na estrutura e na semântica dos dados podem comprometer a eficácia observada nos benchmarks. Para mitigar esse problema, o estudo aponta a necessidade de maior investimento na curadoria de bases de dados específicas e no refinamento de técnicas como calibração de viés e otimização de prompts, tornando os modelos mais robustos para aplicações práticas.

Enquanto isso, as pesquisas de Bartczak [20], Henriques [21] e Maj et. al [22] se concentram nos desafios da implementação de modelos RAG em contextos reais, abordando principalmente o aumento da complexidade e o custo computacional. Henriques [21] propõe uma abordagem de otimização baseada na recuperação de contexto, demonstrando que essa técnica pode melhorar a performance do modelo, embora o ganho em precisão seja ainda limitado. Sua pesquisa destaca como o uso adequado do contexto pode ajudar a reduzir custos computacionais. Por outro lado, Maj et. al [22] argumenta que o excesso de contexto pode prejudicar a precisão das consultas, sugerindo que ele age como

uma forma de overfitting, comprometendo a qualidade das respostas geradas. Essa ambiguidade nos resultados evidencia que a aplicação de RAG para consultas SQL ainda é um campo em evolução, sem diretrizes consolidadas sobre a melhor forma de balancear contexto e desempenho. Como se trata de uma área de pesquisa recente, os próximos passos para aprimorar esses modelos ainda não estão definidos, exigindo mais estudos para estabelecer abordagens mais eficazes e replicáveis.

3.3 QUAIS DESAFIOS TÉCNICOS TÊM SIDO RELATADOS NA UTILIZAÇÃO DE MODELOS DE RAG PARA CONSULTAS SQL?

Essa pergunta busca compreender os obstáculos técnicos, o que é crucial para orientar melhorias futuras. Embora os Modelos de Geração Aumentada por Recuperação tenham mostrado grande potencial na construção automatizada de consultas SQL, desafios técnicos têm sido relatados na literatura. Essas dificuldades estão relacionadas, principalmente, à adaptação dos modelos aos contextos reais de uso, à precisão das consultas geradas, e ao custo computacional envolvido

O principal desafio identificado nas pesquisas está na adaptação dos modelos ao ambiente real. Orrenius [13] observa que, embora os modelos RAG apresentem boa performance em ambientes controlados e em benchmarks como o Spider, eles ainda enfrentam dificuldades em se adaptar a dados reais e complexos. Isso ocorre porque, em cenários do mundo real, a estrutura dos dados pode variar, o que exige ajustes finos e o uso de técnicas de transferência de aprendizado para melhorar a eficácia das consultas geradas. A falta de dados rotulados específicos e a necessidade de treinamento com dados reais tornam essa adaptação um processo demorado e custoso.

Outro desafio relatado é o aumento da complexidade computacional. Bartczak [20], Henriques [21] e Maj et. al [22] apontam que a utilização de RAG para gerar consultas SQL pode levar a um aumento no custo computacional, especialmente quando se trabalha com grandes volumes de dados e consultas complexas. Isso ocorre porque os modelos RAG dependem de mecanismos de recuperação de contexto, que podem envolver um grande número de parâmetros e interações com fontes externas de dados, o que exige mais recursos computacionais para processamento. Isso, atualmente, ainda é um grande impeditivo para o uso em tempo real dessas soluções [20].

A precisão das consultas também tem se mostrado uma preocupação na área.

Henriques [21] discute como o uso de contexto na geração de consultas pode melhorar a performance, mas observa que o ganho em precisão ainda é limitado, especialmente em consultas mais complexas. A dificuldade em obter uma resposta precisa está relacionada ao fato de que o modelo pode falhar em interpretar corretamente a intenção do usuário, especialmente quando há ambiguidade na linguagem natural ou em requisitos específicos dos dados. A pesquisa de Henriques [21] juntamente analisada com a pesquisa de Maj et. al [22] mostra que há uma linha tênue a ser trabalhada, entre o aumento de contexto melhorando performance e o uso limitado de contexto para garantir precisão, reduzindo alucinações.

Por fim, o desafio da interpretação de consultas complexas também é abordado por alguns estudos. Syrjä [15], Rossi [16] e Rizzi [17] destacam que, embora os modelos RAG consigam lidar com consultas simples, há dificuldades em processar consultas SQL mais complexas que envolvem múltiplas junções, subconsultas ou expressões avançadas. Isso se deve ao fato de que esses modelos precisam compreender não apenas a estrutura da consulta, mas também a lógica envolvida nas interações com o banco de dados. Syrjä [15] aponta também que a presença de erros estruturais nos bancos de dados, como colunas mal definidas ou relações inconsistentes, pode comprometer ainda mais a geração das consultas, levando a respostas incorretas ou consultas malformadas.

Esses desafios técnicos apontam para a necessidade de um aperfeiçoamento contínuo dos modelos de RAG, a fim de lidar com a complexidade crescente das consultas SQL e garantir que sua aplicação seja eficiente e escalável em ambientes reais.

3.4 COMO MODELOS DE RAG IMPACTAM A OTIMIZAÇÃO E O DESEMPENHO DE CONSULTAS SQL EM COMPARAÇÃO COM ABORDAGENS TRADICIONAIS?

Essa pergunta tem como motivação investigar como diretamente RAG aprimora LLMs text-to-SQL. Os Modelos de Geração Aumentada por Recuperação (RAG) trazem avanços na otimização e no desempenho de consultas SQL, destacando-se principalmente pela automação e flexibilidade. Em comparação com abordagens tradicionais, que exigem escrita manual e complexa de SQL, os modelos RAG oferecem uma maneira mais eficiente de gerar consultas de forma rápida e acessível, segundo Rossi [16]. No entanto, enquanto a automação das consultas melhora a agilidade e facilita a interação, ela não

elimina completamente a necessidade de um controle detalhado, especialmente quando se trata de otimizar consultas complexas ou de alta performance. RAG pode, em alguns casos, ter dificuldades ao lidar com consultas que exigem uma grande precisão em suas operações, como junções complexas ou subconsultas avançadas. A adaptação dos modelos ao ambiente real, mencionada por Orrenius [13], ainda enfrenta desafios para manter a consistência de desempenho em contextos de dados específicos e variáveis.

Em relação à performance, a utilização de RAG pode melhorar a relevância das respostas ao aplicar contextos recuperados, o que facilita a geração de consultas mais precisas, além da possibilidade de trazer referências às informações que estão sendo dadas, aumentando a transparência e confiabilidade [21]. No entanto, esse processo pode aumentar o custo computacional, como indicam Bartczak [20] e Henriques [21], ao envolver mais etapas e maior uso de recursos, o que pode impactar a velocidade de execução, especialmente em bancos de dados de grande escala. Por outro lado, abordagens tradicionais ainda são preferidas, hoje em dia, em ambientes que priorizam controle rigoroso sobre o desempenho, devido à sua previsibilidade e o uso de técnicas de otimização bem estabelecidas, como índices e particionamento de dados. Esses métodos ainda são mais eficientes em contextos em que a consulta precisa ser otimizada e personalizada para o banco de dados específico, o que não é sempre totalmente alcançado pelos modelos LLM, que são mais generalistas.

As pesquisas que quantificam os impactos da utilização de RAG na geração de SQL, como a de Orrenius [13], mostram que a precisão das consultas pode ser aumentada em até 8,5% em comparação com abordagens sem recuperação de contexto. Em consultas mais complexas, que envolvem múltiplas junções e subconsultas, a taxa de acerto subiu de 67% para 81%, evidenciando a capacidade do modelo de lidar melhor com cenários avançados. Além disso, houve uma redução expressiva na geração de consultas SQL malformadas, que caiu de 12,4% para 3,8%, tornando o uso de RAG uma alternativa viável para minimizar erros sintáticos. Em resumo, enquanto os modelos RAG oferecem melhorias notáveis na geração e relevância das consultas SQL, principalmente ao permitir que usuários sem conhecimento técnico interajam com dados complexos, as abordagens tradicionais ainda se destacam pela capacidade de otimizar consultas em termos de performance e controle técnico em cenários exigentes.

4 AMEAÇAS À VALIDADE

As ameaças à validade representam um desafio constante em revisões sistemáticas da literatura (RSLs), especialmente na área de tecnologia e inovação, nas quais as mudanças e evoluções são constantes. Diante disso, faz-se necessário avaliar as principais ameaças a este estudo, considerando fatores que podem impactar a abrangência, precisão e relevância dos resultados obtidos.

Uma das principais limitações deste trabalho está na temporalidade do tema. Como RAG e LLMs aplicados à construção de consultas SQL são assuntos emergentes, a maioria dos artigos analisados foi publicada recentemente, concentrando-se no ano de 2024. Esse aspecto reforça a atualidade e a relevância do estudo, mas também indica que o campo ainda está em desenvolvimento e pode sofrer mudanças nos próximos anos, afinal, segundo Wohlin [23], revisões podem se tornar rapidamente desatualizadas em áreas dinâmicas.. Assim, a veloz evolução do tema pode tornar esta RSL desatualizada, exigindo revisões contínuas à medida que novas pesquisas são publicadas para mitigar essa ameaça.

Outro fator a ser considerado é o viés de seleção, decorrente da escolha das bases de dados utilizadas. Este estudo se limitou ao Google Scholar, o que pode ter restringido o acesso a trabalhos relevantes presentes em outras bases indexadoras. Ainda que o Google Scholar agregue uma ampla variedade de publicações, a inclusão de outras bases, como IEEE Xplore e ACM Digital Library, poderia oferecer uma visão mais abrangente do tema. O uso do Google Scholar como base de dados principal foi uma escolha fundamentada na avaliação de um número relevante de artigos sobre o tema, identificados inicialmente em outras fontes e bases de dados. Para garantir que a busca fosse o mais completa possível, foi realizada uma revalidação constante da string de busca utilizada, ajustando-a conforme novos artigos relevantes eram identificados. Essa prática visou aumentar a abrangência da pesquisa e minimizar o risco de omissões de artigos pertinentes, assegurando que a seleção de estudos fosse ampla e representativa do estado atual do tema.

O viés de publicação também deve ser destacado. Foram excluídos estudos que não passaram por revisão por pares ou que não estavam publicados em periódicos renomados, o que pode ter levado à exclusão de pesquisas inovadoras que ainda não foram formalmente avaliadas pela comunidade acadêmica. Essa restrição impacta a diversidade das contribuições analisadas e pode limitar a inclusão de ideias e abordagens emergentes que

ainda não receberam ampla validação. Embora tenha sido utilizado material de literatura cinza, ou seja, artigos não revisados por pares, em outras partes do artigo para ampliar a compreensão do tema e capturar inovações ainda não publicadas em periódicos revisados por pares, a escolha de não incluir esses estudos na revisão sistemática de literatura foi tomada com o objetivo de assegurar que o protocolo PRISMA fosse seguido corretamente. Tanto o viés de seleção como o viés de publicação foram analisados por Kitchenham e Charters [24].

Além disso, a falta de reprodutibilidade e a subjetividade na análise dos artigos representam ameaças à validade do estudo. A reprodutibilidade é essencial para garantir que outras pesquisas possam replicar os métodos utilizados e obter resultados semelhantes. Para mitigar esse risco, foi seguido o protocolo PRISMA [14], garantindo transparência no processo de seleção e análise dos artigos. No entanto, eventuais ambiguidades nos critérios de inclusão e exclusão podem dificultar a replicação exata desta RSL. Da mesma forma, mesmo com critérios bem definidos, a interpretação de aspectos metodológicos e a categorização dos estudos podem variar. A Tabela 5 descreve cada um desses viéses.

Tabela 5: Ameaças à Validade

Ameaça	**Descrição**
Temporalidade	O rápido avanço do tema pode tornar a revisão sistemática rapidamente desatualizada, pois novas pesquisas e avanços tecnológicos podem surgir após a coleta e análise dos dados.
Viés de Seleção	A limitação nas bases de dados escolhidas pode restringir o escopo da pesquisa, excluindo estudos relevantes que estejam disponíveis em outras fontes ou formatos não considerados.
Viés de Publicação	A exclusão de estudos não publicados em periódicos renomados pode gerar um viés, pois resultados negativos ou menos impactantes podem não ter sido divulgados amplamente, comprometendo a representatividade dos achados.
Falta de Reprodutibilidade	Se o protocolo adotado não permitir fácil replicação da pesquisa, há um risco de que os mesmos resultados não possam ser obtidos novamente, reduzindo a confiabilidade da revisão sistemática.

5 CONCLUSÃO E TRABALHOS FUTUROS

O estudo realizou uma revisão sistemática da literatura com o objetivo de entender a utilização de RAG (Retrieval-Augmented Generation) e Large Language Models (LLMs) na construção de consultas SQL. Esses modelos demonstram potencial para aprimorar a geração e otimização de consultas, porém ainda apresentam desafios, especialmente no que diz respeito ao equilíbrio entre melhoria no processamento [21] e manutenção da precisão dos resultados esperados [22]. A evolução desses modelos é fundamental para garantir que avanços no desempenho não comprometam a qualidade e confiabilidade das respostas geradas.

A revisão sistemática permitiu identificar as principais abordagens adotadas na literatura para a aplicação de RAG e LLMs na geração de consultas SQL, bem como outros desafios mais recorrentes nesse campo. Entre eles, destacam-se a necessidade de aprimoramento na interpretação de consultas complexas [20], a adaptação a diferentes estruturas de bancos de dados [15] e a mitigação de erros decorrentes de inferências equivocadas dos modelos.

É essencial que pesquisas futuras continuem explorando o aprimoramento dos RAGs e LLMs na construção de consultas SQL. Ademais, a compreensão das limitações identificadas nesta revisão sistemática reforça a necessidade de desenvolvimento de métodos que tornem esses modelos mais interpretáveis e confiáveis, permitindo sua aplicação em cenários reais com maior segurança e eficiência.

Para trabalhos futuros, recomenda-se expandir a busca de artigos para incluir publicações em outros idiomas, como português e espanhol, e explorar bases adicionais, como IEEE Xplore e ACM Digital Library, para garantir uma visão mais abrangente sobre o tema. Além disso, estudos podem se aprofundar na análise comparativa entre diferentes abordagens de RAGs e LLMs, investigando métricas específicas de desempenho e aplicabilidade prática em sistemas de bancos de dados variados.

REFERÊNCIAS

[1] ZHAO, W. X. et al. A survey of large language models. *ArXiv*, abs/2303.18223, 2023. Available at: <https://api.semanticscholar.org/CorpusID:257900969>.

[2] LI, J. et al. 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*, v. 36, 2024.

[3] LYU, Y. et al. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, ACM New York, NY, 2024.

[4] SHAHADE, A. K.; DESHMUKH, P. V. Enhancing natural language processing: A comprehensive review of retrieval augmented generation. In: IEEE. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*. [S.l.], 2024. p. 609–611.

[5] GUIDE, P. E. *Recuperação Aumentada por Geração*. 2025. Acesso em: 5 fev. 2025. Available at: <https://www.promptingguide.ai/research/rag>.

[6] DENG, N.; CHEN, Y.; ZHANG, Y. Recent advances in text-to-sql: A survey of what we have and what we expect. *International Conference on Computational Linguistics*, abs/2208.10099, 2022. Available at: <https://api.semanticscholar.org/CorpusID:251719280>.

[7] GUO, J. et al. Towards complex text-to-sql in cross-domain database with intermediate representation. *Annual Meeting of the Association for Computational Linguistics*, abs/1905.08205, 2019. Available at: <https://api.semanticscholar.org/CorpusID:159041042>.

[8] JEON, J.-B.; LEE, S.; CHOI, S. A systematic review of research on speech-recognition chatbots for language learning: Implications for future directions in the era of large language models. *Interact. Learn. Environ.*, v. 32, p. 4613–4631, 2023. Available at: <https://api.semanticscholar.org/CorpusID:258545509>.

[9] YE, X. et al. Satisfiability-aided language models using declarative prompting. *Neural Information Processing Systems*, abs/2305.09656, 2023. Available at: <https://api.semanticscholar.org/CorpusID:258715073>.

[10] YANG, Y. et al. Advancing structured query processing in retrieval-augmented generation with generative semantic integration. *Frontiers in Computing and Intelligent Systems*, 2024. Available at: <https://api.semanticscholar.org/CorpusID:273572729>.

[11] ZILETTI, A.; D'AMBROSI, L. Retrieval augmented text-to-sql generation for epidemiological question answering using electronic health records. In: *Clinical Natural Language Processing Workshop*. [s.n.], 2024. Available at: <https://api.semanticscholar.org/CorpusID:268385157>.

[12] ALLU, U.; AHMED, B.; TRIPATHI, V. Beyond extraction: Contextualising tabular data for efficient summarisation by language models. *ArXiv*, abs/2401.02333, 2024. Available at: <https://api.semanticscholar.org/CorpusID:266755981>.

[13] ORRENIUS, A. *Enhancing Text-to-SQL Applications with Retrieval Augmented Generation: How does academic advancements in Text-to-SQL translate to industry usage?* 2024.

[14] MOHER, D. et al. Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *PLoS Med*, v. 6, n. 7, p. e1000097, 2009. Available at: <https://doi.org/10.1371/journal.pmed.1000097>.

[15] SYRJÄ, S.-M. Retrieval-augmented generation utilizing sql database case: web sport statistics application. 2024.

[16] ROSSI, A. Large language models to query your data: Retrieving ads industry users data using natural language.

[17] RIZZI, F. *Developing an Enterprise Chatbot using Machine Learning Models: A RAG and NLP based approach*. Tese (Doutorado) — Politecnico di Torino, 2024.

[18] CHIRAS, M. *Exploration of Different Large Language Models for Retrieval-Augmented Generation in Analyzing Wearable Running Data for Sports Physiotherapy*. Dissertação (B.S. thesis) — University of Twente, 2024.

[19] ERIKSSON, A. *Chat-based Search for Intellectual Property Data*. 2024.

[20] BARTCZAK, Z. *From RAG to Riches: Evaluating the Benefits of Retrieval-Augmented Generation in SQL Database Querying*. 2024.

[21] HENRIQUES, P. D. S. *Augmenting Large Language Models with Context Retrieval*. Dissertação (Mestrado), 2024.

[22] MAJ, M. et al. Optimizing customer support using text2sql to query natural language databases. University of Piraeus. International Strategic Management Association, 2024.

[23] WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2014. (EASE '14). ISBN 9781450324762. Available at: <https://doi.org/10.1145/2601248.2601268>.

[24] KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. v. 2, 01 2007.