

**UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENFERMAGEM
PROGRAMA DE PÓS-GRADUAÇÃO EM
GESTÃO DE SERVIÇOS DE SAÚDE
MESTRADO PROFISSIONAL EM GESTÃO DE SERVIÇOS DE SAÚDE**

Fabíola de Oliveira Simões

**APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL COMO
FERRAMENTA PARA SUPORTE À GESTÃO FARMACÊUTICA EM AMBIENTE
HOSPITALAR**

Belo Horizonte
2026

Fabíola de Oliveira Simões

**APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL COMO
FERRAMENTA PARA SUPORTE À GESTÃO FARMACÊUTICA EM AMBIENTE
HOSPITALAR**

Trabalho de Conclusão apresentado ao curso de Mestrado Profissional do Programa de Pós-graduação em Gestão de Serviços de Saúde, da Escola de Enfermagem da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Mestre em Gestão de Serviços de Saúde.

Área de concentração: Gestão de Serviços de Saúde.

Linha de pesquisa: Tecnologias Gerenciais em Saúde

Orientador: Prof. Dr. Raphael Augusto Teixeira de Aguiar

Coorientador: Prof. Dr. Leonardo Chaves Dutra da Rocha

Belo Horizonte
2026

Simões, Fabíola de Oliveira.

S593a Aplicação de processamento de linguagem natural como ferramenta para suporte à gestão farmacêutica em ambiente hospitalar [recurso eletrônico]. / Fabíola de Oliveira Simões. - - Belo Horizonte: 2026.
112f.: il.
Formato: PDF.
Requisitos do Sistema: Adobe Digital Editions.

Orientador (a): Raphael Augusto Teixeira de Aguiar.
Coorientador (a): Leonardo Chaves Dutra da Rocha.
Área de concentração: Gestão de Serviços de Saúde.
Dissertação (mestrado): Universidade Federal de Minas Gerais, Escola de Enfermagem.

1. Processamento de Linguagem Natural. 2. Sistemas Inteligentes. 3. Serviço de Farmácia Hospitalar. 4. Inteligência Artificial. 5. Dissertação Acadêmica. I. Aguiar, Raphael Augusto Teixeira de. II. Rocha, Leonardo Chaves Dutra da. III. Universidade Federal de Minas Gerais, Escola de Enfermagem. IV. Título.

NLM: W 26.55.A7

Bibliotecário responsável: Fabian Rodrigo dos Santos CRB-6/2697



UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENFERMAGEM
COLEGIADO DE CURSO PÓS-GRADUAÇÃO GESTÃO SERVIÇOS DE SAÚDE

ATA DE NÚMERO 138 (CENTO E TRINTA E OITO) DA SESSÃO PÚBLICA DE ARGUIÇÃO E DEFESA DA DISSERTAÇÃO APRESENTADA PELA CANDIDATA FABIÓLA DE OLIVEIRA SIMÕES PARA OBTENÇÃO DO TÍTULO DE MESTRE EM GESTÃO DE SERVIÇOS DE SAÚDE.

Aos doze dias do mês de fevereiro de dois mil e vinte e seis, às 09:30 (nove horas e trinta minutos), realizou-se, por videoconferência, a sessão pública para apresentação e defesa da dissertação **"APLICAÇÃO DE PROCESSAMENTO DE LINGUAGEM NATURAL COMO FERRAMENTA PARA SUPORTE À GESTÃO FARMACÊUTICA EM AMBIENTE HOSPITALAR"**, da aluna *Fabiola de Oliveira Simões*, candidata ao título de "Mestre em Gestão de Serviços de Saúde", linha de pesquisa "Tecnologias Gerenciais em Saúde". A Comissão Examinadora foi constituída pelos seguintes professores doutores: Raphael Augusto Teixeira de Aguiar, Leonardo Chaves Dutra da Rocha, Maria do Carmo e Mônica Viegas Andrade, sob a presidência do primeiro. Abrindo a sessão, o presidente, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação do seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa da candidata. Logo após, os membros da Comissão se reuniram sem a presença da candidata e do público, para julgamento e expedição do seguinte resultado final:

- ☒ APROVADA;
☐ REPROVADA.

O resultado final foi comunicado publicamente à candidata pelo orientador.

Belo Horizonte, 12 de fevereiro de 2026.

Prof. Dr. Raphael Augusto Teixeira de Aguiar
Membro Titular – Orientador (UFMG)

Prof. Dr. Leonardo Chaves Dutra da Rocha
Membro Titular – Coorientador (UFSJ)

Prof. Dr. Maria do Carmo
Membro Titular (FELUMA)

Profª. Drª. Mônica Viegas Andrade

Membro Titular (UFMG)



Documento assinado eletronicamente por **Raphael Augusto Teixeira de Aguiar, Presidente**, em 13/02/2026, às 17:11, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Leonardo Chaves Dutra da Rocha, Usuário Externo**, em 19/02/2026, às 14:14, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maria do Carmo, Usuário Externo**, em 20/02/2026, às 10:20, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Monica Viegas Andrade, Professora do Magistério Superior**, em 23/02/2026, às 14:00, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador 4954835 e o código CRC 58D08C3A.

AGRADECIMENTO

Agradeço primeiramente a Deus, por guiar meus passos, fortalecer minha perseverança e iluminar meu caminho ao longo desta jornada.

À minha família: minha mãe, Elaine, pelo amor e apoio constante; ao meu esposo, Rodrigo, pelo incentivo e companheirismo; e aos meus irmãos, Daniel e Gabriela, por vibrarem comigo e celebrarem cada conquista. Vocês são meu alicerce, meu porto seguro, e sou profundamente grata por cada gesto de cuidado e incentivo.

Aos meus amigos pelas palavras de confiança e apoio, muito obrigada por estarem ao meu lado.

Ao meu orientador, Raphael Aguiar, pela orientação ao longo desta dissertação e pelo apoio na definição dos membros da equipe de TI, cuja contribuição foi fundamental para a realização deste trabalho; e ao meu coorientador, Leonardo Rocha, pelo direcionamento criterioso e pelas observações precisas que enriqueceram o desenvolvimento deste estudo.

À equipe de trabalho — Antônio, Márcio, Matheus e Thiago — pelo suporte técnico e disponibilidade, garantindo que todos os aspectos computacionais do estudo fossem realizados com excelência.

À diretoria do Hospital Metropolitano Doutor Célio de Castro — Cristina Peixoto, Bruno Belezia, Juliana Pantuza e Mauro Heleno — e à equipe de TI-HMDCC, em especial Bruno Souza e Marco Túlio, pelo apoio institucional.

À minha querida e para sempre Equipe de Farmácia pelo carinho e torcida constantes.

À Aline, Jú Pantuza, Matheus, Pati Alves, Pati Serafim, Rô e Orrara pela colaboração e generosidade em compartilhar conhecimentos essenciais para o desenvolvimento deste trabalho.

Por fim, a todos que, de alguma forma, contribuíram para a concretização deste trabalho, expresso minha sincera e profunda gratidão.

Este estudo foi financiado com recursos do Centro de Inovação e Inteligência Artificial para Saúde (CI-IA Saúde), em parte com recursos da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) Processo nº 2020/09866-4, da Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) Processo nº PPE-00030-21 e da UNIMED Belo Horizonte.

“[...] precisamos ser muito mais centrados no ser humano. Se você olhar onde chegamos na IA, eu diria que o grande triunfo seja o reconhecimento de padrões: muito centrada em tarefas, mas carece de consciência contextual e falta aquela aprendizagem flexível que os humanos possuem. Queremos fazer uma tecnologia que melhore a vida dos seres humanos, torne o nosso mundo mais seguro, nossas vidas melhores e mais produtivas. Tudo isso requer uma camada humana de comunicação e de colaboração.” (Fei-Fei Li, cientista da computação e pioneira em Inteligência Artificial)

RESUMO

A gestão de farmácias hospitalares envolve volumes significativos de dados e requer acesso rápido a informações confiáveis para apoiar os processos de tomada de decisão. Este estudo metodológico teve como objetivo explorar a aplicação do Processamento de Linguagem Natural (PLN) na gestão de farmácias hospitalares por meio do desenvolvimento de um protótipo de *chatbot* baseado em um Modelo Amplo de Linguagem (LLM) da família *Qwen*. A pesquisa foi conduzida em duas fases complementares: a primeira fase de caráter exploratório, dedicada à identificação das principais solicitações de informação relacionadas à gestão de estoques e uma segunda fase de desenvolvimento e avaliação funcional do protótipo, implementado em Python. A avaliação da ferramenta focou em um conjunto estruturado de perguntas representativas das necessidades comuns dos profissionais, resultantes da análise das respostas dos participantes durante a primeira fase. Os resultados mostraram uma taxa de sucesso geral de execução de 85%, indicando que o modelo foi capaz de gerar consultas SQL válidas na maioria dos casos. O desempenho mostrou-se satisfatório para perguntas simples de baixa a moderada complexidade, que produziram respostas consistentes e acionáveis. No entanto, uma proporção significativa de consultas mais complexas não levou a respostas semanticamente corretas, apesar da execução técnica bem-sucedida, evidenciando principalmente limitações na interpretação de regras de negócio implícitas. De modo geral, o protótipo demonstrou viabilidade técnica e potencial para aplicação prática, ao mesmo tempo que ressaltou a necessidade de ajustes metodológicos e refinamentos do modelo para melhorar a precisão das respostas em cenários de tomada de decisão mais complexos.

Palavras-chave: Processamento de Linguagem Natural; Sistemas Inteligentes; Serviço de Farmácia Hospitalar; Inteligência Artificial.

ABSTRACT

Hospital pharmacy management involves significant volumes of data and requires rapid access to reliable information to support decision-making processes. This methodological study aimed to explore the application of Natural Language Processing (NLP) in hospital pharmacy management through the development of a chatbot prototype based on a Large Language Model (LLM) from the Qwen family. The research was conducted in two complementary phases: the first, exploratory phase, dedicated to identifying the main information requests related to inventory management, and a second phase of development and functional evaluation of the prototype, implemented in Python. The evaluation of the tool focused on a structured set of questions representative of the common needs of professionals, resulting from the analysis of participants' responses during the first phase. The results showed an overall execution success rate of 85%, indicating that the model was able to generate valid SQL queries in most cases. Performance proved satisfactory for simple questions of low to moderate complexity, which produced consistent and actionable responses. However, a significant proportion of more complex queries did not lead to semantically correct answers, despite successful technical execution, mainly highlighting limitations in the interpretation of implicit business rules. Overall, the prototype demonstrated technical feasibility and potential for practical application, while also emphasizing the need for methodological adjustments and model refinements to improve the accuracy of responses in more complex decision-making scenarios.

Keywords: Natural Language Processing; Intelligent Systems; Pharmacy Service, Hospital; Artificial Intelligence.

LISTA DE ILUSTRAÇÕES

Figura 1 – Esquema da Assistência Farmacêutica no Âmbito Hospitalar	26
Figura 2 – Exemplos de aplicações da IA na Saúde	33
Figura 3 – Exemplo da tarefa Text-to-SQL.....	41
Quadro 1 – Exemplos de métricas para avaliação de modelos comumente usadas em Deep Learning	42
Figura 4 – <i>Prompt</i> de sistema estruturado utilizado pela equipe de TI para orientar a geração de consultas SQL no protótipo	48
Quadro 2 – Conjunto inicial de perguntas para o protótipo do <i>chatbot</i>	53
Figura 5 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 4)	56
Figura 6 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 8)	57
Figura 7 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 1)	58
Figura 8 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 2)	59
Figura 9 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 11)	60
Figura 10 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 3)	61
Figura 11– Resultado da consulta SQL gerada pelo protótipo (Pergunta 6)	62
Figura 12 – Desempenho do protótipo por consulta.....	65

LISTA DE ABREVIATURAS E SIGLAS

AF	Assistência Farmacêutica
BD	Banco de Dados
CAAE	Certificado de Apresentação para Apreciação Ética
CEP	Comitê de ética em Pesquisa
CFF	Conselho Federal de Farmácia
DL	Deep Learning (Aprendizado profundo)
IA	Inteligência Artificial
HMDCC	Hospital Metropolitano Doutor Célio de Castro
LGPD	Lei Geral de Proteção de Dados
LLM	Large Language Models (Modelos Amplos de Linguagem)
ML	Machine Learning (Aprendizado de máquina)
PBH	Prefeitura de Belo Horizonte
PE	Prompt Engineering (Engenharia de <i>Prompt</i>)
PLN	Processamento de Linguagem Natural
PPP	Parceria público privada
RAG	Recuperação Aumentada por Geração
SBRAFH	Sociedade Brasileira de Farmácia Hospitalar
SQL	Structured Query Language (Linguagem de Consulta Estruturada)
SGBD	Sistema de Gerenciamento de Bancos de Dados
SSA	Serviço Social Autônomo
SUS	Sistema Único de Saúde
TI	Tecnologia da Informação
TCLE	Termo de Consentimento Livre e Esclarecido
TCM	Trabalho de Conclusão de Mestrado
TCUD	Termo de Compromisso de Utilização de Dados

SUMÁRIO

1 INTRODUÇÃO	15
1.1 Trajetória profissional e contextualização do tema no cenário de prática	15
1.2 Breve histórico do projeto, seu financiamento e percalços.....	16
2 OBJETIVOS	18
2.1 Objetivo geral	18
2.2 Objetivos específicos	18
3 JUSTIFICATIVA	19
4 REFERENCIAL TEÓRICO	23
4.1 Parcerias público-privadas no setor hospitalar: perspectiva internacional e contexto brasileiro no âmbito do SUS	23
4.2 Farmácia hospitalar e gestão	24
4.2.1 Desafios na gestão da farmácia hospitalar.....	26
4.3 A farmácia hospitalar no contexto das PPP	28
4.4 Conceitos básicos em Tecnologia da Informação e aplicação na saúde	29
4.4.1 Conceitos	29
4.4.2 Aplicações da IA na saúde	32
4.5 Modelos Amplos de Linguagem (LLM).....	34
4.5.1 Definição e importância dos LLM	34
4.5.2 Modelos <i>Open Source</i>	36
4.6 Aplicações de LLM na saúde	37
4.7 Recuperação Aumentada por Geração	39
4.8 Engenharia de <i>prompt</i> e <i>text-to-SQL</i>	40
4.9 Avaliação dos modelos.....	41
5 METODOLOGIA.....	43
5.1 Fase de pesquisa e planejamento.....	43
5.1.1 Coleta de dados	43

5.1.2 Análise dos dados	44
5.2 Desenvolvimento tecnológico do protótipo	45
5.2.1 Extração e preparação dos dados.....	45
5.2.2 Modelo de linguagem utilizado	46
5.2.3 Uso de Recuperação Aumentada por Geração.....	47
5.2.4 Estratégia de <i>prompt</i>	47
5.2.5 Avaliação técnica do protótipo.....	48
5.3 Considerações éticas	48
6 RESULTADOS E DISCUSSÃO.....	50
6.1 Fase de pesquisa e planejamento.....	50
6.1.1 Dificuldades operacionais no acesso à informação.....	50
6.1.2 Limitações dos sistemas informatizados atuais.....	51
6.1.3 Limitações dos sistemas informatizados atuais.....	51
6.1.4 Expectativas quanto ao uso de um <i>chatbot</i> no apoio à gestão farmacêutica hospitalar.....	51
6.2 Definição das perguntas do protótipo	52
6.3 Desenvolvimento do protótipo	54
6.3.1 Resultados globais do protótipo e análise geral	55
6.3.2 Análise por grupo de perguntas	55
6.3.2.1 Consultas de estoque e disponibilidade imediata.....	55
6.3.2.2 Consultas relacionadas à validade de medicamentos e materiais	58
6.3.2.3 Consultas de consumo e histórico temporal.....	61
6.3.2.4 Consultas de apoio à decisão e gestão estratégica	63
6.3.2.5 Consultas com erro de execução	63
6.3.3 Síntese integrada dos resultados e discussão	63
6.4 Limitações e desafios enfrentados no desenvolvimento do protótipo	66
6.5 Perspectivas futuras	67

7 CONCLUSÃO.....	69
REFERÊNCIAS.....	71
APÊNDICE A – Estrutura do Formulário (Google Forms – Fase Exploratória – Avaliação das Demandas)	78
APÊNDICE B – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)	81
APÊNDICE C – TERMO DE COMPROMISSO DE UTILIZAÇÃO DE DADOS (TCUD)	85
APÊNDICE D – CARACTERIZAÇÃO DA AMOSTRA (FASE DE PESQUISA E PLANEJAMENTO)	88
ANEXO – PRODUTO TÉCNICO.....	89

1 INTRODUÇÃO

1.1 Trajetória profissional e contextualização do tema no cenário de prática

Minha trajetória na Gestão de Serviços de Saúde começou em 2017, quando ingressei na Residência Multiprofissional em Farmácia em Intensivismo da Santa Casa de Belo Horizonte, experiência que me permitiu acompanhar de perto a atuação do farmacêutico hospitalar. Em 2019, fui contratada como supervisora de farmácia, atuando em setores como transplante, maternidade, bloco cirúrgico e ambulatório, com responsabilidades em gestão de insumos, farmácia clínica, supervisão de indicadores, auditorias ISO 9001 e participação em comissões. Em 2021, passei a coordenar a farmácia de uma empresa de Home Care, liderando atividades hospitalares e clínicas em filiais de Minas Gerais e fora do estado, além de participar de auditorias, avaliação de indicadores, capacitação de equipes e assumir a responsabilidade técnica perante o Conselho Regional de Farmácia de Minas Gerais e a Vigilância Sanitária.

Em agosto de 2022, iniciei minhas atividades profissionais no Hospital Metropolitano Doutor Célio de Castro (HMDCC), instituição na qual atuo até o momento. O HMDCC caracteriza-se como um Serviço Social Autônomo (SSA) de direito privado, sem fins lucrativos e detentor do título de filantropia, constituído a partir de uma parceria público-privada (PPP) entre a Prefeitura de Belo Horizonte e a Opy Health Metropolitano. Dentre as minhas atribuições enquanto parte da Equipe de Farmácia SSA-HMDCC, destaco o monitoramento da média de consumo de insumos, com vistas a subsidiar a elaboração de ações direcionadas à redução de riscos de perdas, bem como a interlocução diária com empresas parceiras, atuando como referência para demandas relacionadas aos processos de gestão de estoques.

Há pouco mais de um ano, deixei de integrar a equipe de Farmácia do HMDCC, passando a exercer a função de coordenadora do Núcleo de Ensino, Pesquisa e Extensão da mesma instituição. Essa mudança ampliou de forma significativa minha atuação na gestão em saúde, possibilitando uma visão mais transversal dos processos institucionais, das demandas assistenciais, acadêmicas e administrativas, bem como dos desafios relacionados à integração entre setores, produção de conhecimento e tomada de decisão baseada em dados.

No exercício dessa função, pude compreender, de maneira ainda mais aprofundada e com maior clareza do que quando ingressei no programa de mestrado, a relevância e a necessidade de ferramentas que otimizem o acesso à informação de forma estruturada, ágil e confiável. Enquanto gestora, tornou-se evidente que a disponibilidade de informações qualificadas é elemento central para subsidiar decisões mais assertivas, promover eficiência nos processos, fortalecer a gestão do cuidado e contribuir para a sustentabilidade institucional. Nesse sentido, a temática deste trabalho mantém-se diretamente relacionada à minha trajetória profissional e às demandas concretas do cenário de prática, reafirmando sua pertinência e aplicabilidade no contexto da gestão em serviços de saúde.

1.2 Breve histórico do projeto, seu financiamento e percalços

No processo seletivo para ingresso no Programa de Mestrado Profissional em Gestão de Serviços de Saúde da UFMG, é obrigatório apresentar um pré-projeto contendo uma proposta de intervenção ou tema a ser estudado no ambiente de trabalho. Desde o início, eu sabia que queria explorar o uso da inteligência artificial (IA) no contexto da gestão farmacêutica hospitalar, embora ainda não tivesse clareza sobre como viabilizar essa ideia, considerando que não possuo conhecimentos aprofundados em programação ou Tecnologia da Informação (TI).

Após as primeiras reuniões com meu orientador recém definido, Raphael, conseguimos delinear a proposta de desenvolver um *chatbot* como ferramenta de apoio à gestão farmacêutica hospitalar. Considerando o contexto do projeto, formulamos a hipótese de que seria possível adaptar um modelo amplo de linguagem (LLM) em português para interpretar indagações em linguagem natural sobre estoques e demais dados da farmácia hospitalar, retornando respostas claras, coerentes e operacionalmente úteis, apoiando a gestão de insumos farmacêuticos. A partir dessa ideia, defini a pergunta norteadora do estudo: “Como um protótipo de chatbot baseado em Processamento de Linguagem Natural e Modelos Amplos de Linguagem pode ser estruturado e adaptado para organizar, disponibilizar e responder, em linguagem natural, às demandas informacionais da gestão de insumos farmacêuticos hospitalares?”

Com o objetivo de viabilizar a execução do estudo, em maio de 2024 submetemos o projeto ao Edital Ampliado CI-IA Saúde da UFMG – Chamada de março de 2024, promovido pelo Centro de Inovação em Inteligência Artificial para a

Saúde da UFMG (CI-IA Saúde), ainda antes da qualificação do meu Trabalho de Conclusão de Mestrado (TCM). O projeto foi aprovado pela Câmara Departamental do Departamento de Medicina Preventiva e Social da Faculdade de Medicina da UFMG em 08 de agosto de 2024, data em que também foi submetido ao Comitê de Ética em Pesquisa da UFMG (CEP/UFMG). O parecer final de aprovação do CEP/UFMG foi emitido em 20 de março de 2025.

Embora a divulgação do resultado do edital tenha ocorrido em 23 de setembro de 2024, confirmando a aprovação do projeto para financiamento de aproximadamente R\$ 80.000,00, a liberação efetiva dos recursos financeiros só foi realizada em agosto de 2025. Esse inesperado atraso comprometeu o início das atividades dependentes de recursos financeiros, em especial o recrutamento da equipe de TI, essencial para o desenvolvimento colaborativo do protótipo. Como consequência, o cronograma originalmente planejado para doze meses precisou ser condensado em aproximadamente quatro meses, restringindo a execução completa do conjunto de perguntas inicialmente previsto. Apesar dessa limitação, o projeto terá continuidade após a banca final do mestrado, com as etapas restantes sendo atualizadas no repositório GitHub, garantindo a finalização das atividades não concluídas.

2 OBJETIVOS

2.1 Objetivo geral

Desenvolver um protótipo de *chatbot*, baseado em Processamento de Linguagem Natural, capaz de responder às perguntas cotidianas de farmacêuticos atuantes na gestão de farmácia hospitalar em um hospital de grande porte.

2.2 Objetivos específicos

- Identificar as principais demandas informacionais dos gestores e profissionais da farmácia hospitalar;
- Elaborar um conjunto de dados estruturado contendo contextos de conversação frequentes em perguntas da farmácia hospitalar;
- Desenvolver e adaptar o protótipo de *chatbot* utilizando técnicas de Processamento de Linguagem Natural;
- Testar e validar o protótipo em cenários simulados, avaliando a usabilidade, eficiência e precisão das respostas em situações cotidianas da gestão farmacêutica.
- Documentar o processo e preparar o protótipo para futuras expansões;
- Propor estratégias de continuidade e aprimoramento do protótipo.

3 JUSTIFICATIVA

A qualidade do serviço em ambientes hospitalares é um fator essencial para garantir o bem-estar dos pacientes e a eficiência operacional das instituições. A integração de equipes multidisciplinares e a adoção de tecnologias emergentes, conforme será detalhado no referencial teórico, têm se mostrado fundamentais para alcançar esse objetivo.

O desenvolvimento tecnológico contínuo, aliado à inovação, propicia a criação de ferramentas que transformam a maneira como os serviços de saúde são prestados, justificando a busca por soluções baseadas em inteligência artificial para apoiar a gestão hospitalar.

A inteligência artificial se destaca como ferramenta para gerenciar processos em tempo real, especialmente na gestão de estoques e indicadores hospitalares. A capacidade de processar grandes volumes de dados e fornecer insights em tempo real pode melhorar significativamente a eficiência operacional, permitindo monitorar e otimizar o uso de recursos, prever demandas e evitar desperdícios (Amorim; Barone; Mansur, 2008). Além disso, a IA pode identificar padrões e anomalias nos dados, permitindo que os gestores de saúde tomem decisões informadas e proativas (Jin *et al.*, 2024).

Em um cenário de acelerado avanço tecnológico, compreender os desafios da gestão da farmácia hospitalar leva a reflexão sobre como a inteligência artificial pode oferecer soluções inovadoras. Diante da complexidade das operações logísticas necessárias para a execução eficiente da assistência farmacêutica, é fundamental buscar abordagens que otimizem cada etapa do processo. A aplicação da IA pode proporcionar aos farmacêuticos gestores percepções valiosas ao analisar grandes volumes de dados relacionados a cada uma das etapas da Assistência Farmacêutica (AF). Além disso, a IA pode contribuir para a identificação de padrões, previsão de demanda e detecção de anomalias, permitindo uma gestão mais eficaz e ágil. Ao integrar a inteligência artificial com as práticas tradicionais da gestão farmacêutica hospitalar, é possível alcançar melhorias significativas na qualidade dos serviços prestados e na eficiência operacional.

No contexto específico da gestão farmacêutica hospitalar, a inteligência artificial pode ser explorada para apoiar processos de quantificação e ajuste de parâmetros,

analisando volumes de dados sobre uso de medicamentos e oferecendo recomendações baseadas em evidências (Joshi, 2023).

Em 2019, o gasto federal com medicamentos no Brasil foi de R\$19,8 bilhões. Esse valor representa um aumento de quase 10% em relação a 2018 e mais do que o dobro do gasto em 2008 (Silva, 2020). Tais dados indicam que uma gestão ineficiente de insumos farmacêuticos, além de comprometer a segurança do paciente, pode resultar em grandes perdas financeiras.

Entretanto, até o momento, pouco se avançou no uso específico da IA como auxiliar do profissional farmacêutico (Reis, 2023). Embora esses modelos tenham demonstrado grande potencial em diversas áreas, sua aplicação na gestão farmacêutica ainda é uma área com grandes necessidades de desenvolvimento. À medida que a tecnologia continua a evoluir e a infraestrutura necessária se torna mais acessível, há uma expectativa crescente de que esses modelos possam, em breve, desempenhar um papel mais substancial no apoio à tomada de decisões e otimização de processos contribuindo para uma gestão farmacêutica mais eficiente e precisa (Nascimento Neto *et al.*, 2020).

Conforme apresentado na contextualização deste estudo, o hospital em que a pesquisa foi desenvolvida caracteriza-se como um SSA em parceria público-privada, no qual as fases do ciclo da assistência farmacêutica são divididas entre diferentes atores institucionais. Essa configuração gera desafios específicos no acesso à informação, uma vez que os processos não estão concentrados em uma mesma equipe. O acompanhamento de indicadores e a gestão de estoques dependem de múltiplos relatórios e planilhas manuais, agravados pela limitação do sistema eletrônico, que ainda não permite rastreabilidade completa por lote e validade. O resultado prático dessa configuração é que os profissionais que atuam diretamente ou indiretamente com as atividades relacionadas à gestão de estoque na farmácia hospitalar, incluindo farmacêuticos, coordenação de farmácia e a direção da instituição, precisam investir tempo significativo na localização e consolidação de informações que, em um ambiente mais integrado, deveriam estar disponíveis de forma imediata.

Nesse contexto, as tecnologias baseadas em Processamento de Linguagem Natural e em Modelos Amplos de Linguagem, neste caso, mais precisamente um protótipo de *chatbot*, se apresentam como alternativa promissora. A capacidade desses modelos de interpretar perguntas formuladas em linguagem natural e

convertê-las automaticamente em consultas estruturadas a bancos de dados relacionais oferece uma via concreta para aproximar o usuário não técnico dos dados armazenados nos sistemas hospitalares, sem exigir domínio da linguagem de programação ou conhecimento da estrutura do banco de dados subjacente (Yang et al., 2025).

A pertinência desta proposta não reside apenas em sua viabilidade técnica, mas na correspondência entre o problema que ela endereça e as necessidades identificadas empiricamente junto aos profissionais da instituição estudada. As categorias temáticas emergentes da fase de pesquisa deste estudo não são produto de uma insatisfação genérica com a tecnologia disponível. Refletem, antes, a experiência cotidiana de profissionais que operam em um ambiente no qual a complexidade organizacional da PPP amplifica os já conhecidos desafios da gestão farmacêutica hospitalar, tornando ainda mais urgente o desenvolvimento de ferramentas que apoiem a tomada de decisão de forma estruturada e acessível. Nesse sentido, o protótipo de *chatbot* desenvolvido neste estudo deve ser compreendido não como solução tecnológica isolada, mas como resposta a um conjunto articulado de lacunas que o modelo de PPP, por suas características, tende a tornar mais salientes.

Desta forma, o produto desta pesquisa consiste no protótipo de *chatbot*. Foi elaborado um relatório técnico contendo as especificações da ferramenta, que poderá ser analisado pela diretoria do hospital, com vistas ao seu aprimoramento e posterior implementação em ambiente de produção. A aposta na adoção de um LLM de código aberto operando sobre dados reais extraídos do sistema hospitalar institucional materializa a possibilidade de construir uma ferramenta que seja, ao mesmo tempo, tecnicamente viável, operacionalmente relevante e potencialmente replicável em outros contextos hospitalares de natureza similar. Adicionalmente, a utilização de um modelo de código aberto contribui para a redução dos custos associados ao desenvolvimento e à implementação da solução, favorecendo sua viabilidade técnica e ampliando as possibilidades de reprodutibilidade da ferramenta em diferentes instituições de saúde. Além disso, o código-fonte do protótipo foi disponibilizado em um repositório público na plataforma GitHub,¹ de modo a garantir transparência

¹ GitHub é uma plataforma online que permite armazenar, gerenciar e compartilhar códigos e projetos de software, facilitando a colaboração entre desenvolvedores. Baseado no sistema de controle de versão *Git*, ele registra o histórico de alterações realizadas em arquivos, possibilitando que várias

metodológica e possibilitar sua reprodução e evolução por outros pesquisadores e profissionais interessados.

O desenvolvimento deste trabalho contou com o envolvimento da diretoria do Hospital Metropolitano Doutor Célio de Castro desde a fase de planejamento, tendo a instituição apoiado formalmente a execução do estudo. O relatório técnico elaborado será entregue à diretoria, contendo as especificações do protótipo, os resultados da avaliação e as recomendações para refinamentos futuros, de modo a subsidiar a análise institucional quanto à pertinência e à viabilidade de sua evolução para ambiente de produção.

A transição do protótipo para implantação efetiva no ambiente operacional pressupõe etapas que transcendem o escopo do presente estudo, entre as quais se destaca a realização de um estudo de viabilidade econômica, que deverá considerar os custos associados à infraestrutura tecnológica, à manutenção do sistema e à capacitação das equipes, confrontando-os com os ganhos operacionais esperados em termos de eficiência, redução de perdas e qualificação da tomada de decisão. O HMDCC, instituição que tem a inovação como valor organizacional e histórico de abertura à incorporação de tecnologias que contribuam para a melhoria de seus processos, reúne condições institucionais favoráveis para que essa avaliação ocorra, tornando concreta a possibilidade de adoção da ferramenta, condicionada à demonstração de sua viabilidade em escala de produção.

4 REFERENCIAL TEÓRICO

4.1 Parcerias público-privadas no setor hospitalar: perspectiva internacional e contexto brasileiro no âmbito do SUS

A relação entre o Estado e o setor privado na gestão de serviços hospitalares não é fenômeno recente. Suas raízes modernas remontam às reformas de cunho gerencialista que marcaram a política pública de diversos países a partir dos anos 1980, quando a pressão por eficiência, contenção de gastos e ampliação da oferta de serviços impulsionou a busca por modelos alternativos à provisão estatal direta (Almeida, 2017). Nesse contexto, as parcerias público-privadas (PPP) emergem como arranjo institucional que procura combinar a responsabilidade do Estado pela prestação do serviço com a capacidade operacional, financeira e gerencial do setor privado, sem configurar privatização (Torchia; Calabrò; Morner, 2015).

No plano internacional, Nikolic e Maikisch (2006), em documento de referência do Banco Mundial, identificaram quatro grandes modalidades de PPP na saúde: infraestrutura, prestação de serviços, proteção financeira e modelos híbridos. A experiência europeia constitui o cenário mais documentado. No Reino Unido, as Private Finance Initiatives (PFI), introduzidas a partir dos anos 1990, tornaram-se a principal fonte de investimento de capital no sistema nacional de saúde (NHS), com contratos de arrendamento que cobrem desde a construção de hospitais até serviços de apoio operacional (Almeida, 2017). A Itália, com o Servizio Sanitario Nazionale, é o segundo maior mercado europeu de PPP hospitalar, tendo adotado amplamente o modelo britânico a partir de 2002. Na Espanha, as PPP seguem predominantemente a lógica das concessões administrativas, com processos de licitação estruturados em múltiplas etapas (Torchia; Calabrò; Morner, 2015). Em Portugal, a experiência acumulada com os chamados hospitais PPP revelou resultados heterogêneos: embora 75% das unidades analisadas tenham sido consideradas eficientes em pelo menos um dos anos avaliados, os indicadores não foram uniformemente superiores aos de hospitais de gestão pública, sinalizando que o modelo por si só não é garantia de desempenho (Nunes; Matos, 2017).

Na América Latina, a difusão das PPP hospitalares ocorreu de forma mais gradual e com contornos próprios. No Brasil, a base legal para as PPP foi estabelecida pela Lei Federal nº 11.079, de 30 de dezembro de 2004, que instituiu normas gerais para licitação e contratação de parcerias no âmbito da Administração Pública,

prevendo tanto a concessão patrocinada quanto a concessão administrativa, sendo que nesta a Administração Pública é usuária direta dos serviços, como é o caso dos hospitais públicos. Esta última modalidade passou a ser o fundamento jurídico sobre o qual se estruturaram as primeiras experiências hospitalares no país (Brasil, 2004b).

A inauguração do Hospital do Subúrbio (HS), em Salvador (Bahia), em setembro de 2010, marcou o início efetivo das PPP hospitalares no SUS. Trata-se da primeira unidade pública do país a operar integralmente sob esse modelo, com a concessionária Prodal Saúde responsável tanto pela gestão predial e administrativa quanto pelos serviços clínicos assistenciais, em um contrato de concessão que contempla o período de dez anos. Outros estados brasileiros avançaram na mesma direção nos anos seguintes, com destaque para Minas Gerais, onde o Hospital Metropolitano Doutor Célio de Castro, em Belo Horizonte, passou a operar sob o regime de Serviço Social Autônomo, em parceria público-privada firmada entre a Prefeitura de Belo Horizonte e a Opy Health Metropolitano.

Estudos que avaliaram comparativamente o desempenho de hospitais geridos por PPP no SUS apontam que os indicadores assistenciais (como por exemplo, tempo médio de permanência, taxa de ocupação de leitos e taxa de mortalidade hospitalar) não se mostraram sistematicamente superiores aos de hospitais com outros modelos de gestão, inclusive no caso do HMDCC (Fernandes; Bordin, 2022). Esses achados são consistentes com as evidências internacionais, que apontam para a necessidade de governança contratual robusta, mecanismos efetivos de monitoramento e capacidade regulatória do poder público como condições determinantes para que a parceria gere os resultados esperados (Torchia; Calabrò; Morner, 2015; Almeida, 2017). A simples transferência da gestão ao parceiro privado, desacompanhada de instrumentos de fiscalização e de alinhamento de incentivos, tende a reproduzir ou mesmo amplificar as fragilidades estruturais preexistentes.

4.2 Farmácia hospitalar e gestão

O setor de farmácia hospitalar é responsável por tarefas como seleção, programação e aquisição de materiais médicos e medicamentos para utilização nas diferentes áreas assistenciais do hospital, orientando o uso racional e avaliando a relação custo-benefício. No Brasil, durante a década de 1950, os serviços de farmácia hospitalar, representados pelas Santas Casas de Misericórdia e hospitais-escola, iniciaram um processo de desenvolvimento e modernização. Em 1973, a Lei nº 5.991

determinou que toda farmácia, incluindo a farmácia hospitalar, deveria contar com a presença de um farmacêutico responsável técnico (Conselho Regional de Farmácia do Estado de São Paulo, 2019). Este é o profissional detentor das ferramentas necessárias para um gerenciamento eficiente dos insumos farmacêuticos, podendo resultar em uma redução de custos e em uma maior eficiência nas atividades clínico-assistenciais realizadas em uma instituição (Melo; Oliveira, 2021).

No ambiente hospitalar, a farmácia, sob a gestão de um profissional farmacêutico, tem como propósito fornecer suporte abrangente, garantindo a adequação dos insumos e sua integração eficaz com as operações hospitalares (Packer; Resta, 2014). Segundo o Conselho Federal de Farmácia (CFF), a farmácia hospitalar e outros serviços de saúde são definidos como uma “unidade clínica, administrativa e econômica, liderada por um farmacêutico, que está hierarquicamente ligada à direção do hospital ou serviço de saúde e integrada funcionalmente com as demais unidades administrativas e de assistência ao paciente” (Conselho Regional de Farmácia do Estado de São Paulo, 2019). As etapas da AF devem ser pensadas de forma cíclica. Todas influenciam e são influenciadas umas pelas outras (Marin *et al.*, 2003; Brasil, 2007).

As atividades realizadas pela farmácia hospitalar podem ser analisadas considerando a perspectiva da organização sistêmica da AF. Conforme definido pela Resolução nº 338/2004 do Conselho Nacional de Saúde, a Assistência Farmacêutica é:

[...] um conjunto de ações voltadas à promoção, proteção e recuperação da saúde, tanto individual como coletiva, tendo o medicamento como insumo essencial e visando ao acesso e ao seu uso racional. Este conjunto envolve a pesquisa, o desenvolvimento e a produção de medicamentos e insumos, bem como a sua seleção, programação, aquisição, distribuição, dispensação, garantia da qualidade dos produtos e serviços, acompanhamento e avaliação de sua utilização, na perspectiva da obtenção de resultados concretos e da melhoria da qualidade de vida da população (Brasil, 2004a, p.1).

A integração entre a farmácia hospitalar e a Assistência Farmacêutica garante um fluxo contínuo de medicamentos e insumos hospitalares de qualidade, promovendo uma abordagem multidisciplinar para atender às necessidades dos pacientes com eficácia. No ambiente hospitalar, a AF engloba diversas atividades como logística, manipulação, garantia da qualidade, cuidados farmacêuticos e farmácia clínica, como ilustrado no esquema da Figura 1. Além disso, existem atividades que requerem colaboração entre departamentos, envolvendo a interação com outras áreas do hospital.

Figura 1 – Esquema da Assistência Farmacêutica no Âmbito Hospitalar



Fonte: Adaptado de Conselho Regional de Farmácia do Estado de São Paulo (2019).

As atribuições do farmacêutico hospitalar no Brasil são definidas pela Resolução do CFF nº 568/2012, que regulamenta o exercício profissional nos serviços de atendimento pré-hospitalar, na farmácia hospitalar e em outros serviços de saúde, de natureza pública ou privada. Dentre as atribuições definidas para este profissional, é importante destacar:

Art. 4º – São atribuições do farmacêutico nos serviços de atendimento pré-hospitalar, farmácia hospitalar e outros serviços de saúde:

III. Gestão da informação, infraestrutura física e tecnológica;

Art. 5º – Nas atividades de assistência farmacêutica, é de competência do farmacêutico nos serviços de atendimento pré-hospitalar, farmácia hospitalar e outros serviços de saúde:

I. Assumir a coordenação técnica nas ações relacionadas à padronização, programação, seleção e aquisição de medicamentos, insumos, matérias-primas, produtos para saúde e saneantes, buscando a qualidade e a otimização da terapia medicamentosa (Conselho Federal de Farmácia, 2012, p. 2-3).

4.2.1 Desafios na gestão da farmácia hospitalar

A Portaria nº 4.283, de 30 de dezembro de 2010, estabelece as diretrizes para o fortalecimento da farmácia hospitalar no Brasil. A AF é fortemente dependente de operações logísticas para sua execução, abrangendo todo o Ciclo da Assistência Farmacêutica. No entanto, a interface desse ciclo está fundamentada nas ações de atenção à saúde centradas no usuário, ultrapassando as questões estritamente logísticas (Bruns; Luiza; Oliveira, 2014). O objetivo da gestão na farmácia hospitalar é

garantir o abastecimento, dispensação, acesso, controle, rastreabilidade e uso racional de medicamentos, além de otimizar a relação entre custo, benefício e risco (Brasil, 2010).

Espera-se que a farmácia hospitalar desenvolva as atividades clínicas e relacionadas à gestão, que devem ser organizadas de acordo com as características da instituição onde o serviço é prestado, mantendo assim a consistência com o tipo e nível de complexidade do hospital. Parte das dificuldades enfrentadas está associada à escassez de recursos financeiros, que impactam em aspectos como recursos humanos, materiais e logística (Raimundo; Dias; Guerra, 2015; Lima; Câmara, 2016).

Manter o equilíbrio econômico-financeiro é um desafio constante para os gestores hospitalares, especialmente diante da necessidade de otimização dos reduzidos recursos financeiros, atendendo ao aumento da demanda por serviços de saúde e garantindo a oferta adequada de produtos para a saúde sem que isso impacte na qualidade da assistência prestada ao paciente (Lima; Câmara, 2016). Uma gestão farmacêutica com processos organizacionais bem estruturados, direcionados à melhoria contínua, desempenha um papel estratégico na superação dos desafios enfrentados, independentemente da natureza pública ou privada da instituição. Isso se deve ao fato de que produtividade e qualidade estão intimamente interligadas e são elementos essenciais para alcançar resultados positivos (Barbosa, 2015).

A complexidade e rapidez das transformações no mundo globalizado destacam a importância primordial do desenvolvimento e fortalecimento da gestão para garantir a sustentabilidade das organizações. Portanto, é essencial ampliar os estudos sobre a gestão da assistência farmacêutica e destacar, em termos práticos, a importância de se consolidarem metodologias que possam ser eficazes na análise dos dados disponíveis e no apoio à gestão (Bruns; Luiza; Oliveira, 2014).

Estudo realizado por Lima e Câmara (2016) evidenciou que a inadequação da estrutura física e a insuficiência de recursos humanos qualificados também são problemas importantes enfrentados na gestão da cadeia de suprimentos em um hospital. O primeiro representa um desafio significativo na gestão da cadeia de suprimentos hospitalares, impactando diretamente a eficiência da armazenagem e distribuição dos produtos de saúde. Essa deficiência pode resultar em dificuldades operacionais, atrasos e até mesmo desperdícios de recursos. Além disso, a insuficiência de recursos humanos qualificados, como farmacêuticos e assistentes, agrava ainda mais os problemas enfrentados. A escassez de profissionais

capacitados para atender à demanda do estabelecimento de saúde não apenas compromete a qualidade dos serviços prestados, mas também aumenta os riscos de desabastecimento e falhas na gestão dos medicamentos e insumos essenciais.

Por fim, a dificuldade no acesso e utilização adequada de tecnologias pode representar um desafio significativo para a gestão farmacêutica hospitalar, influenciando diretamente a disponibilidade de produtos para a saúde dentro da instituição. A ausência de ferramentas tecnológicas adequadas pode resultar em processos manuais, lentos e propensos a erros, dificultando o monitoramento preciso dos estoques, a previsão de demanda e a gestão de pedidos (Lima; Câmara, 2016).

4.3 A farmácia hospitalar no contexto das PPP

A inserção da farmácia hospitalar em instituições geridas sob o modelo de PPP impõe configurações organizacionais que diferem substancialmente das encontradas em hospitais de gestão pública direta. Em um hospital público convencional, a cadeia da AF tende a ser coordenada por uma única estrutura técnico-administrativa, ainda que sujeita às limitações orçamentárias e às rigidezes da burocracia pública. No contexto das PPP hospitalares, a responsabilidade pelas atividades relacionadas aos insumos farmacêuticos e hospitalares é compartilhada entre os entes público e privado. No HMDCC, contexto em que o presente estudo está inserido, as etapas de seleção, programação, aquisição, prescrição e monitoramento permanecem sob responsabilidade do parceiro público, enquanto as atividades de armazenamento, distribuição e dispensação são executadas pelo parceiro privado, conforme estabelecido no modelo de gestão adotado pela instituição.

Entre os aspectos potencialmente favoráveis desse arranjo, destaca-se a possibilidade de incorporação de práticas gerenciais do setor privado à gestão dos estoques farmacêuticos. Contudo, o funcionamento da farmácia hospitalar em ambientes de parceria público-privada introduz desafios que vão além dos já conhecidos na literatura sobre gestão farmacêutica em geral. Um primeiro ponto de tensão diz respeito à fragmentação das responsabilidades ao longo do ciclo da AF. Quando diferentes etapas estão distribuídas entre atores institucionais distintos, surgem descontinuidades na cadeia de informações que dificultam o acompanhamento integral do estoque e comprometem a rastreabilidade por lote e validade.

A dificuldade de acesso à informação qualificada em tempo oportuno constitui talvez o desafio mais transversal à gestão farmacêutica em hospitais sob qualquer modelo de gestão, mas assume contornos particulares nas PPP. Quando os processos não estão concentrados em uma única equipe e os dados relevantes se encontram distribuídos entre diferentes bases, o farmacêutico gestor se vê frequentemente dependente de consultas a múltiplas fontes para responder a demandas rotineiras (como a simples verificação do quantitativo em estoque real, o acompanhamento de perdas e baixas do estoque ou a identificação de medicamentos próximos ao vencimento). Esse conjunto de dificuldades aponta para a necessidade de ferramentas que centralizem o acesso à informação, reduzam a dependência de processos manuais e ampliem a capacidade analítica dos profissionais envolvidos na gestão farmacêutica hospitalar.

4.4 Conceitos básicos em Tecnologia da Informação e aplicação na saúde

4.4.1 Conceitos

a) Inteligência Artificial

A IA refere-se à capacidade de um computador executar tarefas comumente associadas a seres humanos como a capacidade de raciocinar, descobrir significados, generalizar ou aprender com a experiência passada (Tacca; Rocha, 2018). Desde a década de 1940, estudos têm demonstrado que os computadores podem realizar tarefas muito complexas, como descobrir provas para teoremas matemáticos ou jogar xadrez, com grande eficácia. A inteligência artificial, impulsionada por modelos de linguagem avançados, está sendo aplicada em diversas áreas, como diagnóstico médico, motores de busca, reconhecimento de voz e *chatbots* (Copeland, 2024).

As áreas-chave da IA incluem visão computacional, aprendizado de máquina, aprendizado profundo, robótica, representação de conhecimento e raciocínio e PLN. A visão computacional foca em ensinar máquinas a interpretar e compreender o mundo visual, processando imagens e vídeos para identificar objetos, rostos, cenas e atividades. O aprendizado profundo, uma subcategoria do aprendizado de máquina, utiliza redes neurais profundas para analisar grandes conjuntos de dados, frequentemente alcançando reconhecimento de padrões de alto nível. A robótica combina diversas disciplinas para construir e programar robôs que possam realizar tarefas variadas, muitas vezes substituindo ou auxiliando humanos em atividades

perigosas, repetitivas ou de alta precisão. A representação de conhecimento e raciocínio estuda como os sistemas podem representar conhecimentos do mundo real e raciocinar sobre eles para tomar decisões lógicas e informadas (Aguilar, 2024a).

b) Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural é uma vertente da inteligência artificial, a qual refere-se ao desenvolvimento de modelos computacionais para compreensão e processamento da linguagem humana em sua forma oral ou escrita. Assim, tecnologias com aplicação do PLN estão evoluindo para permitir que os computadores interajam com os humanos de forma mais natural e intuitiva (Caseli; Freitas; Viola, 2022).

A criação de programas com uso de PLN é complexa, uma vez que, tradicionalmente, os computadores são projetados para que os humanos conversem com eles através de uma linguagem de programação precisa, inequívoca e altamente estruturada, ou usando um número limitado de comandos de voz claramente articulados. Porém, a linguagem humana nem sempre é precisa, é ambígua e a sua estrutura linguística pode depender de diversos fatores, como gírias, dialetos regionais e contexto social.

c) Geração de Linguagem Natural

A geração de linguagem natural é uma tecnologia de inteligência artificial que permite que máquinas criem texto de forma inteligente e automatizada. As primeiras abordagens baseavam-se em regras gramaticais, mas os avanços recentes em Machine Learning (ML) e Deep Learning (DL) revolucionaram o ensino da linguagem humana às máquinas (Dörr, 2015). Dessa forma, a IA é capaz de gerar uma linguagem coerente usando técnicas de previsão da próxima palavra baseadas em probabilidade. Para o sistema, isso significa decidir o que responder e como apresentar essa resposta ao usuário. Esta capacidade enriquece as interações pessoa-máquina e, em particular, a qualidade das respostas em *chatbots* e outras interfaces de conversação. Atualmente, o GPT (Generative Pre-Trainer Transformer), da OpenAI,² é o exemplo de maior sucesso, sendo capaz de gerar língua de forma tão fluente quanto os humanos (Floridi; Chiriatti, 2020).

² <https://chat.openai.com/>

As abordagens tradicionais dependem de regras gramaticais para produzir um texto coerente. Os sistemas que utilizam esse método definem regras específicas para construção de frases e montagem de palavras, o que ajuda a gerar conteúdo estruturado. Esses modelos, muitas vezes baseados em redes neurais, aprendem a compreender a estrutura da linguagem analisando grandes quantidades de dados textuais (Floridi; Chiriatti, 2020).

d) Sistema de Gerenciamento de Banco de Dados

Um sistema de banco de dados (BD) é uma aplicação computadorizada projetada para manter registros e facilitar o armazenamento, recuperação e atualização de informações conforme necessário. As informações armazenadas podem variar amplamente, representando qualquer coisa que tenha relevância para os usuários individuais ou para a organização que utiliza o sistema (Date, 2004). Há uma distinção entre “dados” e “informações”, segundo a qual o primeiro se refere aos registros brutos armazenados no banco de dados, enquanto o segundo se refere ao significado desses dados para os usuários (Sumathi; Esakkirajan, 2007; Rob; Coronel, 2010).

O Sistema de Gerenciamento de Banco de Dados (SGBD) é uma coleção de programas que administram a estrutura do banco de dados e supervisionam o acesso aos dados armazenados (Rob; Coronel, 2010). Portanto, o SGBD tem a responsabilidade de controlar, acessar, organizar e proteger as informações de uma aplicação. Além disso, atua como intermediário na comunicação entre o usuário e os dados armazenados. Por meio de uma linguagem de consulta como o SQL, do inglês, Structured Query Language, é viável inserir, atualizar, gerenciar, recuperar e remover dados do banco.

e) Linguagem de Consulta Estruturada (SQL)

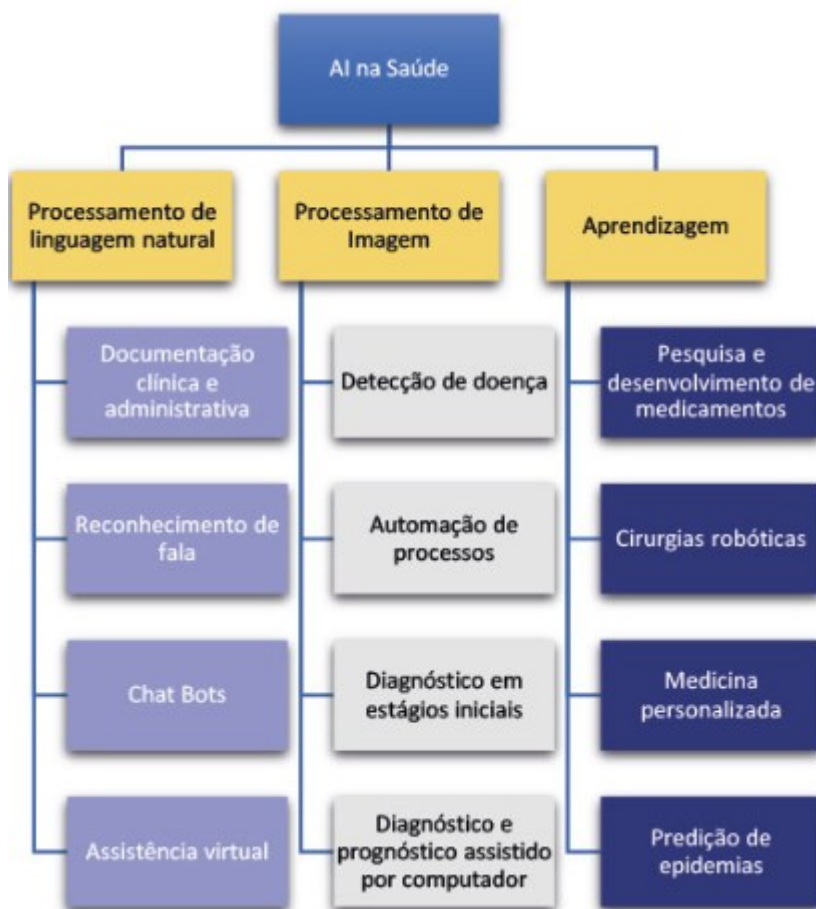
A Linguagem de Consulta Estruturada (SQL) é utilizada para controlar e gerenciar dados em um sistema de banco de dados relacional. Apesar de ser associada principalmente à realização de consultas, a SQL abrange diversas funções como inserção, atualização e exclusão de informações. Os bancos de dados relacionais são estruturados em tabelas compostas por linhas e colunas, permitindo assim representar diferentes características e relações entre os elementos. Além disso, neste modelo teórico são disponibilizados operadores para a manipulação das linhas das tabelas, os quais facilitam o processo de dedução de novas proposições

verdadeiras a partir das proposições fornecidas (Date, 2004). Fundamentada na álgebra relacional, com instruções SQL é possível realizar operações como armazenamento, modificação, remoção, busca e recuperação de informações do BD, além de otimizar seu desempenho.

4.4.2 Aplicações da IA na saúde

Cada vez mais presente no dia a dia, a IA é utilizada em ferramentas de pesquisa como o Google, redes sociais, sistemas de recomendação de *streamings*, carros autônomos e assistentes virtuais com reconhecimento de fala humana, como a Siri e a Alexa. Na saúde, a ciência tem presenciado durante os últimos anos, um avanço notável na aplicação destas ferramentas. A Figura 2 apresenta algumas aplicações na saúde. Embora extensa, a lista apresentada não é exaustiva, já que essa é uma área dinâmica em que as pesquisas estão em pleno desenvolvimento e o campo continua a evoluir rapidamente, trazendo novas possibilidades e benefícios para profissionais e pacientes (Reis, 2023).

Figura 2 – Exemplos de aplicações da IA na Saúde



Fonte: Reis (2023).

Sistemas de IA podem auxiliar no diagnóstico de doenças ao analisar sintomas, detectar interações medicamentosas, exames e histórico do paciente para fornecer diagnósticos mais precisos e rápidos. Eles também podem desenvolver planos de tratamento personalizados baseados em características individuais, monitorar sinais vitais dos pacientes para prevenir complicações, e ajudar na gestão de dados médicos, identificando padrões e tendências. Além disso, a IA contribui para a medicina de precisão, ao identificar biomarcadores e padrões genéticos, e para a análise de imagens médicas, auxiliando na detecção precoce de doenças e na interpretação de resultados (Iliashenko; Bikkulova; Dubgorn, 2019; Nascimento Neto *et al.*, 2020).

Bzhalava *et al.* (2022) afirma que as tecnologias digitais estão focadas na digitalização e automação de diagnósticos de saúde, bem como no desenvolvimento de serviços de saúde remotos ou domiciliares. Além disso, menciona que tecnologias

como inteligência artificial e aprendizado de máquina estão associadas ao cuidado de idosos e ensaios clínicos.

Para Rong *et al.* (2020), as categorias de aplicação de inteligência artificial na área da saúde são: assistência à vida para idosos e pessoas com deficiência, técnicas de Processamento de Linguagem Natural, atividades de pesquisa fundamentais, diagnóstico e previsão de doenças. Essas categorias visam tratar grandes volumes de dados, fornecer acesso rápido a informações e melhorar a qualidade dos cuidados de saúde.

Sendo assim, a IA pode ajudar os profissionais de saúde a identificar sinais que poderiam ser facilmente perdidos, devido à sobrecarga de trabalho e à quantidade de variáveis a considerar. Empresas como a IBM Watson e Google têm desenvolvido soluções em saúde que utilizam a IA e o PLN, como um sistema de perguntas e respostas para apoiar decisões médicas. A IA não apenas fornece orientações, mas também pode detectar predisposições a doenças ou identificá-las precocemente, mesmo quando não são visíveis a olho nu. Existem também aplicativos que utilizam análise facial, IA e *insights* genômicos para melhorar diagnósticos e tratamentos de doenças raras (Iliashenko; Bikkulova; Dubgorn, 2019).

Frente às inúmeras possibilidades de aplicação da IA e do PLN na saúde, é relevante destacar os LLMs, os quais surgem como importante estratégia para o desenvolvimento de soluções avançadas como os *chatbots*. Esses modelos, treinados com vastas quantidades de dados textuais, possuem a capacidade de compreender e gerar linguagem humana de maneira sofisticada. Na área da saúde, eles podem ser utilizados para criar assistentes virtuais que auxiliam no atendimento ao paciente, fornecem informações médicas precisas e personalizadas, e até ajudam profissionais da saúde na interpretação de dados clínicos complexos. A implementação de *chatbots* avançados pode melhorar a eficiência dos serviços de saúde, reduzir a carga de trabalho dos profissionais e proporcionar um atendimento mais ágil e acessível aos pacientes.

4.5 Modelos Amplos de Linguagem (LLM)

4.5.1 Definição e importância dos LLM

Um ramo específico da Inteligência Artificial é o aprendizado de máquina, ou Machine Learning, que utiliza técnicas matemáticas e estatísticas para aprender com

os dados e fazer inferências (Benning; Peintner; Peinter, 2022). As formas mais avançadas de aprendizado de máquina são conhecidas como aprendizado profundo, ou Deep Learning, que envolvem múltiplos níveis de características ou variáveis. Considerada uma evolução do ML, o DL estrutura algoritmos em camadas para criar uma rede neural artificial capaz de aprender e tomar decisões inteligentes de forma autônoma (Schuhmacher *et al.*, 2021).

Deep Learning é uma subárea da IA que utiliza redes neurais artificiais para aprender e executar tarefas. Essas redes têm muitas camadas de unidades de processamento e aprendem a representar dados em escalas abstratas e hierárquicas. O DL aprende automaticamente com base nos dados sem programação explícita e tem sido aplicado com sucesso em Processamento de Linguagem Natural, visão computacional e áreas médicas (Dong; Wang; Abbas, 2021). Destaca-se o sucesso notável em diagnósticos médicos auxiliados por computador.

Em uma rede neural, cada neurônio realiza uma operação essencial, que envolve a aplicação de uma função linear aos dados de entrada, ajustada por valores chamados pesos e bias. Os pesos são valores que determinam a influência das entradas de um neurônio na saída do próximo neurônio. Sua função principal é definir a importância de cada neurônio artificial em relação a um determinado processamento. Inicialmente, esses pesos podem ser atribuídos de maneira igual ou aleatória. Durante o treinamento do modelo, eles são ajustados conforme o desempenho da rede. O bias é um valor que assegura que, mesmo quando todas as entradas são zero, a saída do neurônio ainda pode ser diferente de zero. Ele oferece uma flexibilidade adicional à função de ativação, atuando como um ajuste fino essencial para o aprendizado eficaz. Tanto os pesos quanto os bias são chamados genericamente de parâmetros (Aguiar, 2024b).

Modelos Amplos de Linguagem são a aplicação prática do DL no domínio da linguagem natural, treinados com vastos conjuntos de dados e focados em reconhecer, resumir, traduzir, prever e gerar texto. Essa tecnologia é capaz de aprender a estrutura e estilos linguísticos, capturando nuances e padrões semânticos. Uma de suas características fundamentais é a habilidade de gerar textos coerentes e fluentes, além de poder completar frases, traduzir, contar histórias e responder perguntas. O LLM utiliza uma arquitetura avançada de aprendizado de máquina, baseada em redes neurais profundas. Essa abordagem é essencial para compreender e gerar linguagem natural de forma contextualmente precisa.

Modelos Amplos de Linguagem estão se tornando mais acessíveis e eficientes, embora a obtenção de dados de treinamento suficientes continue sendo um desafio. Ainda assim, mesmo em conjuntos de dados pequenos, técnicas apropriadas podem permitir que esses dados sejam altamente eficazes para diversas tarefas de IA, apesar dos desafios envolvidos.

4.5.2 Modelos *Open Source*

Um LLM *Open Source* (ou de “código aberto”, em português) é um Modelo Amplo de Linguagem cujos componentes centrais como pesos do modelo e arquitetura são disponibilizados publicamente de forma que pesquisadores e desenvolvedores possam baixar, estudar, modificar e executar o modelo em sua própria infraestrutura, sem restrições proprietárias impostas por fornecedores fechados. Essa definição enfatiza a transparência e a possibilidade de reprodução, elementos considerados fundamentais na literatura sobre democratização de modelos de IA (Singer *et al.* 2024).

Entre os LLM *Open Source* mais relevantes na atualidade destacam-se as diferentes variantes da família *LLaMA*, além dos modelos *DeepSeek*, *Kimi K2*, *Mistral*, *Gemma*, as famílias *Qwen* e *Phi*, amplamente empregados em pesquisas e aplicações práticas nas áreas de Processamento de Linguagem Natural e inteligência artificial. É importante ressaltar que o panorama de LLMs evolui rapidamente, com novos modelos surgindo e ganhando destaque em curtos períodos de tempo, de modo que listas específicas tendem a ficar desatualizadas em pouco tempo após sua publicação.³ Por esse motivo, apenas o modelo efetivamente usado neste projeto, o QWEN3-32B-AWQ, será destacado.

³ Diretório atualizado de modelos LLM que agrega milhares de modelos em uma base de dados pesquisável: <https://huggingface.co/models>.

QWEN3-32B-AWQ

O Qwen3-32B-AWQ é um modelo avançado de IA para linguagem, desenvolvido pela Qwen Team. Ele faz parte de uma nova geração de modelos capazes de entender e produzir textos em vários idiomas. Este modelo é o maior de arquitetura densa dentro da família Qwen3, contando com 32 bilhões de parâmetros. Sua arquitetura é baseada no modelo anterior Qwen2.5, utilizando técnicas como Grouped Query Attention (GQA), SwiGLU, Rotary Positional Embeddings (RoPE) e RMSNorm com pré-normalização, além da introdução do QK-Norm para garantir a estabilidade do treinamento (Qwen Team, 2025).

Uma das principais inovações do Qwen3-32B é a integração de dois modos de operação em um único *framework*, sendo capaz de alternar entre um modo de “raciocínio mais detalhado” e um modo de resposta direta, dependendo do tipo de comando enviado. Ele é capaz de realizar tarefas como criar textos criativos, responder perguntas, resolver problemas matemáticos, auxiliar em programação e participar de conversas.

O sufixo AWQ (Activation-aware Weight Quantization) refere-se a uma técnica de quantização de 4 *bits*. Esta técnica é aplicada ao Qwen3-32B para reduzir significativamente o uso de memória VRAM e acelerar a inferência, sem perda substancial de precisão. O AWQ protege os pesos mais importantes do modelo (aqueles que geram ativações maiores) durante o processo de compressão, tornando o Qwen3-32B-AWQ uma escolha ideal para implantação em hardwares com recursos de memória mais limitados, mantendo as capacidades de raciocínio descritas no relatório técnico (Qwen Team, 2025).

Esse tipo de modelo pode ser usado em *chatbots*, ferramentas educacionais, sistemas de criação de conteúdo e assistentes de programação, entre outras aplicações.

4.6 Aplicações de LLM na saúde

Os modelos de linguagem estão sendo empregados na área da saúde para acelerar a criação de diretrizes clínicas, fornecer respostas rápidas e precisas a questões complexas, auxiliar no diagnóstico e tratamento de pacientes, além de gerenciar condições crônicas. Esse uso pode aprimorar a precisão dos diagnósticos, diminuir erros e aumentar a eficácia na prestação de cuidados aos pacientes (Tustumi; Andreollo; Aguilar-Nascimento, 2023).

Modelos de linguagem como o GPT são usados na saúde para prever doenças potenciais, fornecer conselhos de saúde personalizados e melhorar a precisão e eficiência no diagnóstico. Eles são capazes de lidar com a complexidade dos dados médicos, oferecendo uma compreensão profunda dos textos médicos e melhorando a gestão da saúde (Montagna *et al.* 2023; Jin *et al.*, 2024).

Na assistência ao paciente, *chatbots* alimentados por LLMs são empregados para responder a perguntas comuns dos pacientes, oferecer informações sobre sintomas, tratamentos e medicamentos, permitindo que os profissionais de saúde se concentrem em casos mais complexos e no atendimento direto aos pacientes. No desenvolvimento de conteúdo médico, LLMs são usados para redigir documentos clínicos, como prontuários médicos e relatórios de exames, otimizando o tempo dos profissionais de saúde e garantindo registros precisos e completos. Além disso, na análise de dados clínicos, os LLMs podem extrair informações valiosas de grandes conjuntos de dados textuais, como registros médicos eletrônicos e literatura científica, auxiliando na identificação de padrões epidemiológicos, na predição de doenças e na personalização de tratamentos, o que contribui para a tomada de decisões clínicas mais informadas e eficazes (Lee; Bubeck; Petro, 2023).

Montagna *et al.* (2023) apresenta, em seu estudo, o desenvolvimento de *chatbot* programado e treinado para interagir com pacientes hipertensos, motivando-os a medir a pressão arterial, seguir prescrições, monitorar a evolução da pressão e possivelmente gerar alerta se os sinais vitais excederem a faixa segura. O trabalho concluiu que o uso desta tecnologia pode promover o empoderamento dos pacientes hipertensos, ajudando-os a monitorar sua pressão arterial e aderir às prescrições médicas. O *chatbot* interage com os pacientes em linguagem natural, fornecendo lembretes, verificando a precisão dos dados inseridos e alertando sobre valores fora da faixa segura, além de manter conversas empáticas com os pacientes, o que pode melhorar a experiência do usuário e a adesão ao tratamento.

Adaptar-se ao conhecimento em saúde, o qual está em constante evolução, requer a implementação de mecanismos de aprendizado contínuo nos LLMs, permitindo que atualizem periodicamente suas bases de conhecimento a partir de fontes confiáveis. Além disso, é essencial personalizar as informações de saúde, desenvolvendo LLMs com capacidades adaptativas que se ajustem ao histórico do paciente e suas preferências, incorporando *feedbacks* para aprimorar as respostas.

Integrar esses modelos com módulos de inteligência artificial especializados e promover equipes interdisciplinares pode facilitar a interpretação de dados complexos.

4.7 Recuperação Aumentada por Geração

Os LLMs têm sido utilizados com sucesso em diferentes atividades, como tradução de textos, elaboração de resumos e, sobretudo, na resposta a perguntas abertas. Esse tipo de pergunta, em que o usuário procura informações ou esclarecimentos que não estão previamente estruturados, impõe desafios relevantes. Nessa situação, o modelo deve interpretar corretamente a solicitação e produzir uma resposta coerente, precisa e compatível com o contexto apresentado. Dessa forma, os LLMs contribuem para a obtenção de respostas mais objetivas e organizadas, favorecendo a agilidade no acesso às informações (Soares, 2024; Yang *et al.* 2025).

Em contextos específicos, como a gestão farmacêutica hospitalar, a utilização da técnica conhecida como Recuperação Aumentada por Geração (do inglês, Retrieval-Augmented Generation – RAG) tem sido explorada como uma solução potencial para melhorar o desempenho dos LLMs na oferta de informações mais precisas e confiáveis. Essa técnica permite que o sistema consulte previamente bases de dados internas antes de gerar uma resposta. Dessa forma, o modelo não se baseia apenas em conhecimentos gerais, mas utiliza informações específicas do contexto em que está inserido, o que torna as respostas mais adequadas (Soares, 2024; Lewis *et al.*, 2020; Databricks, [2025]; Yang *et al.*, 2025).

Como exemplo, pode-se considerar a situação em que um usuário consulta um modelo de linguagem sobre os protocolos mais recentes para uso de um medicamento em ambiente hospitalar. Modelos tradicionais podem não dispor dessas informações atualizadas, o que pode resultar em respostas desatualizadas ou “alucinadas”. Com a utilização da técnica de RAG, o sistema realiza previamente a busca por documentos atualizados e relevantes e utiliza essas informações no momento da geração da resposta. Dessa forma, o modelo passa a se apoiar em dados atuais e pertinentes para formular uma resposta consistente e alinhada ao contexto hospitalar. Essa abordagem permite que o LLM forneça informações mais confiáveis, mesmo em temas para os quais não foi diretamente treinado, reduzindo o risco de desinformação e contribuindo para a melhoria da qualidade do conteúdo gerado (Hoshi *et al.*, 2023; Soares, 2024).

Segundo Yang *et al.* (2025), a aplicação da RAG no contexto da IA generativa, especialmente na área da saúde, apresenta contribuições relevantes em três aspectos principais: equidade, confiabilidade e personalização. Em especial, no que se refere à personalização, destaca-se o potencial da RAG para apoiar a gestão individualizada do cuidado em saúde. De modo geral, modelos generativos de IA apresentam limitações para incorporar informações pessoais dos usuários, o que pode dificultar a oferta de orientações de saúde mais eficazes. Em contraste, a abordagem baseada em RAG permite a integração de dados personalizados, possibilitando a construção de um perfil mais abrangente. Essa característica pode contribuir para a oferta de recomendações em saúde mais alinhadas às necessidades específicas de cada pessoa.

4.8 Engenharia de *prompt* e *text-to-SQL*

O *prompt* consiste em uma instrução estruturada e contextualizada, destinada a orientar modelos de linguagem generativa na produção de respostas mais adequadas, precisas e pertinentes ao objetivo proposto. Pode ser compreendido como o elemento inicial do processo de interação, responsável por explicitar ao sistema as expectativas quanto ao conteúdo e à forma da resposta esperada (Brasil, 2025).

A engenharia de *prompt* (do inglês Prompt Engineering – PE), por sua vez, refere-se ao campo de estudo voltado à elaboração, ao aprimoramento e à otimização dessas instruções, com a finalidade de potencializar a qualidade, a exatidão e a aplicabilidade das respostas geradas pelos modelos de inteligência artificial (Brasil, 2025; Zhou *et al.*, 2023). Uma vez que LLMs possuem treinamentos complexos, onerosos e genéricos, a PE permite que os modelos tenham melhor desempenho em casos de uso específicos para cada usuário (Feng *et al.*, 2024).

Por meio da PE, torna-se possível especificar previamente o formato no qual a resposta deverá ser apresentada em relação à entrada fornecida. Nesse sentido, é possível utilizar a PE para solicitar que o modelo produza como saída um código de programação em uma linguagem definida pelo usuário ou, ainda, que gere automaticamente consultas em linguagem SQL, por meio da aplicação da técnica *text-to-SQL* (Silva; Silva; Silva, 2025). Em termos práticos, a técnica *text-to-SQL* consiste em permitir que consultas sejam formuladas em linguagem natural (texto), cabendo ao modelo de linguagem interpretar essa entrada e traduzi-la automaticamente em

comandos SQL, de modo a viabilizar a recuperação das informações solicitadas (Figura 3).

Figura 3 – Exemplo da tarefa Text-to-SQL



Fonte: Silva; Silva; Silva (2025).

4.9 Avaliação dos modelos

As métricas de avaliação são medidas fundamentais utilizadas para avaliar a qualidade de modelos. Elas são usadas para determinar o desempenho de classificadores em termos de precisão, separabilidade, calibração e sensibilidade a mudanças na distribuição de classe prévia. Cada família dessas métricas tem características específicas que a tornam adequada para diferentes tipos de problemas (Ferri; Hernández-Orallo; Modroiu, 2009). O Quadro 1 apresenta alguns exemplos das métricas comumente utilizadas em Deep Learning.

Após avaliado adequadamente, o modelo pode sofrer ajustes ou mesmo ser submetido novamente a uma ou mais etapas de sua preparação, visando à melhora de sua performance. Por fim, quando esta for considerada satisfatória, o modelo é disponibilizado para uso “no mundo real” (ou, segundo o jargão técnico, colocado “em produção”), para que faça previsões ou execute outras atividades conforme projetado. O desempenho do modelo em produção deve ser monitorado continuamente e ajustado conforme necessário.

Quadro 1 – Exemplos de métricas para avaliação de modelos comumente usadas em Deep Learning

Métricas de Classificação	Acurácia (<i>Accuracy</i>): A proporção de previsões corretas em relação ao total de previsões.
	Precisão (<i>Precision</i>): A proporção de verdadeiros positivos em relação ao número total de exemplos que foram classificados como positivos.
	Sensibilidade (<i>Recall</i>): A proporção de verdadeiros positivos em relação ao número total de exemplos que realmente são positivos.
	F1 Score: A média harmônica entre a precisão e o <i>recall</i> .
	AUC-ROC (Área sob a Curva ROC): Mede a capacidade do modelo de distinguir entre classes.
Métricas de Regressão	Erro Médio Absoluto (<i>Mean Absolute Error - MAE</i>): A média das diferenças absolutas entre as previsões e os valores reais.
	Erro Quadrático Médio (<i>Mean Squared Error - MSE</i>): A média das diferenças quadráticas entre as previsões e os valores reais.
	Raiz do Erro Quadrático Médio (<i>Root Mean Squared Error - RMSE</i>): A raiz quadrada do MSE, que oferece uma métrica na mesma escala que os valores alvo.
	R² Score (Coeficiente de Determinação): Mede a proporção da variação nos dados de saída que é explicada pelo modelo.
Métricas de Avaliação de Modelos de Clustering em Texto	Pureza (<i>Purity</i>): Mede a extensão em que cada <i>cluster</i> contém apenas pontos de dados de uma única classe.
	Informação Mútua Normalizada (<i>Normalized Mutual Information (NMI)</i>): Avalia a qualidade do <i>clustering</i> comparando a similaridade entre os <i>clusters</i> obtidos e as classes reais.

Fonte: Elaborado pela autora. Baseado em Ferri, Hernández-Orallo e Modroiu (2009).

5 METODOLOGIA

Estudo metodológico para o desenvolvimento de um protótipo de *chatbot*, baseado em Processamento de Linguagem Natural. Na literatura de engenharia de *software*, a prototipagem é definida como prática orientada à exploração e validação de soluções em desenvolvimento, caracterizada por propósito, escopo, mídia, forma de uso e estratégia de exploração bem delimitados (Bjarnason; Lang; Mjöberg, 2023). O propósito pode variar desde a investigação de viabilidade técnica até a avaliação de funcionalidades em cenários representativos do domínio de aplicação, o que corresponde à concepção de protótipo em sentido estrito. Um protótipo é o primeiro exemplar ou modelo de um produto, exibindo suas características técnicas e funcionalidades, e permite aos desenvolvedores testar o uso e identificar possíveis falhas antes da implantação em ambiente de produção (Nair *et al.*, 2023). Na construção de um protótipo de *chatbot*, as etapas envolvem a definição do escopo e objetivos, preparação e seleção de dados, desenvolvimento do sistema e realização de testes para ajustes e melhorias, conforme metodologia proposta por Shevat (2017).

Na literatura de saúde digital e sistemas conversacionais, o termo protótipo é empregado consistentemente para designar sistemas que implementam funcionalidades reais, operam sobre dados representativos do domínio e são submetidos a avaliação quantitativa, mas que ainda não se encontram implantados em contexto operacional pleno (Wilson; Marasoiu, 2022; Nair *et al.*, 2023). Caracterização esta que corresponde integralmente ao produto desenvolvido neste estudo, cujo sistema opera sobre dados reais extraídos do sistema de gestão hospitalar da instituição, foi avaliado por métricas quantitativas formais em comparação com um conjunto de referência validado, e dispõe de interface web para visualização dos resultados, sem configurar, todavia, implantação em ambiente de produção. O estudo foi conduzido em duas fases complementares, detalhadas nas subseções seguintes.

5.1 Fase de pesquisa e planejamento

5.1.1 Coleta de dados

Foi realizada pesquisa exploratória, do tipo descritiva, para discussão, estudo e análise das principais demandas às quais o protótipo de *chatbot* deveria atender. Essa etapa foi desenvolvida com foco nos profissionais direta ou indiretamente

envolvidos nos processos relacionados à farmácia hospitalar, a saber: três farmacêuticos(as) gestores de estoque, um enfermeiro(a) da qualidade de materiais, três farmacêuticos(as) gestores de processos, um(a) coordenador(a) de farmácia, um(a) assessor(a) da diretoria e um(a) diretor(a) da diretoria de recursos terapêuticos, diagnóstico e segurança assistencial. Foram excluídos os profissionais que estavam de férias ou afastados do trabalho por qualquer motivo no período da coleta de dados.

Neste estágio, em virtude dos cargos ocupados pelos participantes e das limitações de disponibilidade na agenda, foi disponibilizado um formulário no Google Forms (Apêndice A), sendo este elaborado com os seguintes agrupamentos: dados de caracterização, conhecimento autorreferido sobre assistência farmacêutica no âmbito hospitalar e principais informações consultadas rotineiramente e que podem amparar o processo de gestão de estoques.

Foram convidados a participar do estudo dez profissionais pertencentes ao quadro funcional da instituição, os quais atendiam aos critérios de inclusão estabelecidos. Todos manifestaram aceite inicial e assinaram o Termo de Consentimento Livre e Esclarecido (Apêndice B). Posteriormente, foi encaminhado por correio eletrônico o *link* de acesso ao formulário eletrônico no Google Forms, com a definição de um prazo para o envio das respostas.

Apesar do envio de lembretes, um dos profissionais convidados não respondeu ao instrumento de coleta de dados, resultando em um total de nove participantes efetivos na pesquisa (n=9). Para fins de análise e apresentação dos resultados, os participantes foram identificados por códigos alfanuméricos, denominados de P1 a P9, de modo a assegurar o anonimato e a confidencialidade das informações.

5.1.2 Análise dos dados

Os dados obtidos por meio do formulário eletrônico foram submetidos a uma abordagem metodológica mista, combinando análise quantitativa descritiva das questões fechadas e análise qualitativa de conteúdo das questões abertas.

As questões fechadas, de caráter sociodemográfico e funcional, foram analisadas por meio de estatística descritiva, utilizando-se medidas de frequência absoluta e relativa, bem como medidas de tendência central (média e mediana).

As questões abertas foram analisadas por meio de análise categorial temática, conforme proposta de Bardin (1977), realizada em três etapas:

- **Pré-análise**, com leitura flutuante das respostas para familiarização com o material;
- **Exploração do material**, com identificação de unidades de sentido e agrupamento das respostas em categorias temáticas;
- **Tratamento e interpretação dos resultados**, com organização das categorias emergentes e descrição dos temas mais recorrentes.

As categorias temáticas foram definidas a partir do próprio conteúdo das respostas.

A análise foi conduzida com foco exploratório, visando orientar a construção de um conjunto inicial de perguntas e intenções de uso que reflete as necessidades reais da equipe farmacêutica.

5.2 Desenvolvimento tecnológico do protótipo

5.2.1 Extração e preparação dos dados

Para viabilizar o desenvolvimento do protótipo, a equipe de Tecnologia da Informação do hospital realizou a extração dos dados diretamente da base institucional. O hospital utiliza o sistema eletrônico de gestão hospitalar Tasy®, cujas informações são armazenadas em um banco de dados Oracle®. A partir desse sistema, foi realizado um recorte específico da base de dados, contemplando exclusivamente as informações relevantes para o escopo do estudo.

Considerando o tempo necessário para a extração dos dados, o volume dos arquivos gerados e a complexidade do processamento, optou-se por trabalhar, inicialmente, apenas com informações relacionadas a materiais médicos e medicamentos. Essa delimitação permitiu a construção de uma base de dados adequada para a etapa inicial de desenvolvimento, treinamento e validação do protótipo, assegurando a viabilidade técnica do projeto e a aderência aos objetivos propostos. A extração incluiu dados referentes ao cadastro de materiais médicos e medicamentos no sistema, bem como os registros de movimentações de estoque associados a esses itens, compreendendo o período de 2022, 2023 e 2024. As informações extraídas foram disponibilizadas pela equipe de TI do HMDCC em arquivos no formato .csv, os quais foram armazenados em um drive compartilhado institucional, garantindo acesso controlado e seguro aos pesquisadores responsáveis pelo estudo.

O projeto foi conduzido em colaboração com uma equipe de TI com conhecimentos avançados em SQL, incluindo a capacidade de elaboração e otimização de consultas complexas, bem como proficiência em Python para manipulação e transformação de dados. A equipe também possuía experiência na implementação de rotinas de Extração, Transformação e Carga e na limpeza e preparação de conjuntos de dados destinados a aplicações de ML. A atuação da equipe teve início em agosto de 2025.

Inicialmente, os dados extraídos do sistema hospitalar foram organizados e estruturados em um banco de dados *PostgreSQL*, criando-se um ambiente controlado para análise, testes e integração com o modelo de linguagem. Paralelamente, foi desenvolvido um dicionário de dados contendo a descrição das tabelas e colunas extraídas, bem como a explicitação das principais regras de negócio associadas às informações de estoque. Esse dicionário teve como finalidade alinhar o entendimento entre os pesquisadores e a equipe de TI do estudo, possibilitando a correta interpretação dos dados e das perguntas definidas na fase de planejamento, além de subsidiar a elaboração das consultas SQL utilizadas no desenvolvimento e na validação do protótipo.

Como etapa fundamental do desenvolvimento, foi construída uma base de referência denominada *ground truth*, composta por consultas SQL elaboradas e validadas pela equipe de TI. O *ground truth* corresponde a um conjunto de consultas consideradas corretas e representativas das perguntas de interesse, servindo como padrão para comparação e avaliação das consultas geradas automaticamente pelo modelo. Essa base é essencial em tarefas de *text-to-SQL*, pois permite verificar se as consultas produzidas pelo modelo recuperam resultados equivalentes aos obtidos pelas consultas de referência (Krantz; Jonker, [2025]).

5.2.2 Modelo de linguagem utilizado

Para o desenvolvimento do protótipo foi selecionado um LLM da família Qwen, especificamente o modelo Qwen3-32B-AWQ, por sua natureza *Open Source*, capacidade de compreensão semântica em língua portuguesa e desempenho satisfatório em tarefas de geração de texto estruturado.

O modelo Qwen foi empregado com a finalidade de interpretar perguntas formuladas em linguagem natural por usuários da área de gestão farmacêutica

hospitalar e traduzi-las em consultas estruturadas na linguagem SQL, compatíveis com a estrutura do banco de dados institucional.

5.2.3 Uso de Recuperação Aumentada por Geração

Conforme conceituado no referencial teórico (seção 4.7), a técnica de Recuperação Aumentada por Geração (RAG) foi utilizada neste estudo. Na gestão farmacêutica hospitalar, os bancos de dados reúnem grande quantidade de informações organizadas com termos próprios da área. Nesse contexto, o RAG foi empregado para auxiliar o sistema a acessar o banco de dados de forma adequada, respeitando a organização e a nomenclatura das informações. Dessa maneira, as perguntas realizadas ao sistema passam a gerar respostas mais claras, corretas e confiáveis.

5.2.4 Estratégia de *prompt*

A estratégia de *prompt* adotada neste estudo refere-se à forma como as instruções foram apresentadas ao modelo de linguagem para orientar o tipo de resposta esperada.

A Figura 4 apresenta o *prompt de sistema* utilizado para orientar o modelo de linguagem na conversão de perguntas em linguagem natural para consultas SQL precisas. Esse *prompt* define de forma explícita o papel do modelo, as diretrizes de geração das consultas, as regras de negócio adotadas para análise de estoque e os critérios de formatação das respostas.

Ressalta-se que o *prompt de usuário* consiste exclusivamente na pergunta em linguagem natural formulada pelo usuário final, enquanto o *prompt* apresentado na figura permanece fixo, assegurando padronização, reprodutibilidade e alinhamento com os objetivos do sistema.

Figura 4 – *Prompt* de sistema estruturado utilizado pela equipe de TI para orientar a geração de consultas SQL no protótipo

```
Você é um conversor de linguagem natural para SQL. Sua tarefa é converter
perguntas em linguagem natural em consultas SQL precisas.

Contexto do Banco de Dados:
{context} é trocado pelas informações do banco, tabelas, colunas, descrições e
os relacionamentos

Regras de negócio:
{business_rules}

Diretrizes Importantes:
- Sempre que possível mostre o id do medicamento/item, adicione o no select.
- Sempre utilize os nomes de tabelas e colunas corretos, conforme mostrado no
contexto.
- Utilize JOINS explícitos ao acessar múltiplas tabelas.
- Utilize cláusulas WHERE apropriadas para filtragem.
- Considere valores NULL em suas consultas.
- Utilize funções de agregação (SUM, COUNT, AVG) quando apropriado.
- Sempre utilize aliases para as tabelas para maior clareza.
- Quando existir operações de única coluna no select, nomeie a coluna com o
nome de antes da operação.
- Palavras como "histórico", "listagem", "registros" ou "movimentações"
indicam que o usuário quer os dados detalhados, sem agregação.
- Toda query com "últimos X tempo" DEVE usar no filtro: BETWEEN (CURRENT_DATE
- intervalo) AND CURRENT_DATE

Regras de Resposta:
1. NÃO inicie com "Aqui está o SQL" ou explicações.
2. NÃO utilize blocos de código markdown (sql). Retorne apenas o texto cru da
query.
3. NÃO inclua comentários ou notas finais.

Output esperado: Apenas a string SQL válida iniciada por SELECT.

Regras de Negocio:
## Definição de estoque crítico: Estoque crítico é o nível mínimo de estoque
abaixo do qual existe risco de desabastecimento. Considera-se estoque crítico
quando a cobertura é igual ou menor que 15 dias de consumo médio.

### Consumo médio mensal:
O consumo médio mensal é calculado considerando os últimos 3 meses.

Como se calcula:
Calcular o consumo médio diário:
CMD=Consumo mensal / 30 dias
Calcular a cobertura em dias:
Cobertura (dias)=Quantidade em estoque disponível / Consumo mensal
Comparar com o valor crítico:
- Se cobertura ≤ 15 dias → Estoque crítico
- Se cobertura > 15 dias → Estoque adequado ou de segurança
```

Fonte: Elaborado e utilizado pela equipe de TI no desenvolvimento do protótipo (2025).

5.2.5 Avaliação técnica do protótipo

O desempenho de cada pergunta do conjunto de avaliação foi analisado por meio de métricas quantitativas para avaliação de sistemas de recuperação de informação e geração automática de consultas, tais como acurácia, precisão, *recall* e *F1-score*. Essas métricas foram calculadas a partir da comparação entre as consultas SQL geradas automaticamente pelo modelo e a base de referência (*ground truth*) previamente validada.

5.3 Considerações éticas

O projeto de pesquisa foi submetido à apreciação do Comitê de Ética em Pesquisa da Universidade Federal de Minas Gerais (CEP/UFMG) e à instituição local onde o estudo foi desenvolvido, tendo sido aprovado conforme as recomendações da

Resolução nº 466/12 do Conselho Nacional de Saúde (Brasil, 2012), sob o número de Certificado de Apresentação para Apreciação Ética (CAAE) 3242424.7.1001.5149.

Os participantes envolvidos na fase de pesquisa e planejamento foram devidamente esclarecidos acerca do Termo de Consentimento Livre e Esclarecido (TCLE), constante no Apêndice B, o qual foi assinado em duas vias antes do início das atividades.

Na fase de desenvolvimento, o estudo utilizou informações provenientes do sistema eletrônico da instituição. Para esse fim, foi apresentado e aprovado o Termo de Compromisso de Utilização de Dados (TCUD), constante no Apêndice C, autorizando a coleta e a utilização dos dados necessários à execução da pesquisa. Os riscos associados à privacidade e à confidencialidade dos dados foram mitigados por meio da adoção de estratégias rigorosas de anonimização, controle de acesso e segurança da informação.

O texto da presente dissertação foi submetido à avaliação por ferramenta de inteligência artificial (*Claude Opus 4.5*), com a finalidade de apoio à revisão linguística, clareza textual e organização do conteúdo. Ressalta-se que a ferramenta não foi utilizada para a elaboração de resultados, interpretação dos dados ou tomada de decisões metodológicas, sendo a responsabilidade intelectual integralmente atribuída à autora.

6 RESULTADOS E DISCUSSÃO

O estudo foi desenvolvido em duas fases complementares. A fase de pesquisa e planejamento, realizada com profissionais que atuam rotineiramente nos processos de gestão farmacêutica hospitalar, contribuiu de forma decisiva para a identificação e a delimitação das perguntas que integrariam o conjunto de consultas contempladas pelo protótipo e empregadas na etapa subsequente de seu desenvolvimento, assegurando o alinhamento com as necessidades dos usuários.

6.1 Fase de pesquisa e planejamento

Os dados provenientes das questões fechadas, por apresentarem menor relevância aos objetivos centrais deste estudo, são apresentados no Apêndice D.

As respostas abertas do questionário foram submetidas à análise de conteúdo, conforme descrito na seção de metodologia. A amostra foi composta por nove participantes, pertencentes a dois eixos profissionais: gerencial (n=3) e operacional assistencial (n=6).

A partir da análise realizada, as respostas foram organizadas nas seguintes categorias temáticas:

1. Dificuldades operacionais no acesso à informação
2. Limitações dos sistemas informatizados atuais
3. Demandas por agilidade, centralização e confiabilidade da informação
4. Expectativas quanto ao uso de um *chatbot* no apoio à gestão farmacêutica hospitalar.

6.1.1 Dificuldades operacionais no acesso à informação

A primeira categoria, relacionada às dificuldades operacionais no acesso à informação, revelou-se transversal aos dois eixos profissionais. No entanto, as respostas indicaram diferenças quanto ao impacto dessas dificuldades no cotidiano de trabalho. Entre os profissionais do eixo operacional assistencial, predominaram relatos sobre a demora para localizar informações, a necessidade de realizar múltiplas consultas e a ausência de dados atualizados em tempo real, aspectos que interferem diretamente na rotina assistencial e logística. Já no eixo gerencial, as dificuldades estiveram mais associadas à obtenção de informações consolidadas, necessárias para o acompanhamento dos processos e para a tomada de decisão.

Exemplo:

“O chatbot poderia deixar o **acesso mais ágil, melhorar a visibilidade das informações e reduzir erros de análise**” (P5).

6.1.2 Limitações dos sistemas informatizados atuais

A segunda categoria diz respeito às limitações dos sistemas informatizados atualmente utilizados. As percepções apresentadas foram semelhantes entre os participantes, independentemente do eixo de atuação no hospital, destacando-se a fragmentação das informações decorrente da falta de integração entre os sistemas. No eixo operacional assistencial, essas limitações repercutem diretamente na fluidez do trabalho cotidiano, enquanto, nos cargos de gestão, comprometem a consolidação dos dados e a análise global dos processos.

Exemplo:

“Carência de um sistema de informações que permita o rastreamento estoquelote de cada medicamento para **validação de validades** e tomada de decisão quanto aos **estoques próximos do vencimento, dificuldade de rastreamento de informações** via e-mail de forma prática e ágil para o conhecimento do **histórico de consumo de um item** e tomada de decisão quanto ao seu **planejamento de compras**, dificuldades de visualização de ordens de fornecimento em atraso ou pendências financeiras relacionadas a fornecedores de **itens que não estão críticos**” (P4).

6.1.3 Limitações dos sistemas informatizados atuais

A terceira categoria evidenciou demandas por maior agilidade, centralização e confiabilidade das informações. Esse tema apareceu de forma recorrente nos discursos, sendo reconhecido como uma necessidade comum aos dois eixos profissionais.

Exemplo:

“Um chatbot pode facilitar bastante o **acesso às informações de estoque** ao oferecer uma interface rápida, intuitiva” (P8).

6.1.4 Expectativas quanto ao uso de um *chatbot* no apoio à gestão farmacêutica hospitalar

A quarta categoria diz respeito às expectativas quanto ao uso de um *chatbot* no apoio à gestão farmacêutica hospitalar. De modo geral, os participantes

compreenderam essa tecnologia como uma possibilidade de responder às fragilidades identificadas nas categorias anteriores. No eixo operacional assistencial, o *chatbot* foi associado à facilitação do acesso às informações, ao suporte às atividades diárias e à redução do tempo dedicado à busca por dados. Já no eixo gerencial, as expectativas estiveram relacionadas ao potencial do *chatbot* para integrar informações, apoiar a tomada de decisão e auxiliar na geração de relatórios gerenciais.

Exemplo:

“Cruzar os dados de **previsão de demanda**, com atrasos de entrega e tempo médio de entrega de cada fornecedor” (P1).

“Ajudaria com informações centralizadas, facilitando a visualização rápida de dados como **validade** ou **quantidade em estoque**.

Rapidez no acesso aos **históricos de consumo e previsões de demanda**, facilitando o planejamento e a reposição.

Dados atualizados sobre fornecedores e prazos de entrega, facilitando negociações ou respostas rápidas a urgências” (P3).

6.2 Definição das perguntas do protótipo

Os resultados obtidos na fase de pesquisa e planejamento possibilitaram a definição das perguntas que compõem o protótipo do *chatbot*. As categorias temáticas identificadas a partir da análise de conteúdo evidenciaram fragilidades no contexto da gestão farmacêutica do hospital. Esses achados orientaram a elaboração de perguntas alinhadas às demandas concretas relatadas pelos participantes, buscando transformar necessidades identificadas no campo empírico em funcionalidades do protótipo.

As dificuldades operacionais no acesso à informação, caracterizadas pela demora na localização de dados, pela necessidade de múltiplas consultas e pela ausência de informações atualizadas em tempo oportuno, justificam a inclusão de perguntas voltadas ao monitoramento imediato do estoque. Questões relacionadas à identificação de itens críticos, à verificação da quantidade disponível e à detecção de situações de falta respondem diretamente às dificuldades descritas pelos profissionais do eixo operacional assistencial, ao mesmo tempo em que subsidiam decisões gerenciais relacionadas ao abastecimento.

A necessidade de acesso ao histórico de consumo e ao controle de lotes e validades evidencia lacunas importantes para o planejamento de compras e para a prevenção de desperdícios. Nesse sentido, as perguntas que respondam a esses questionamentos buscam suprir a dificuldade relatada de obtenção dessas informações de forma prática e integrada.

A categoria relacionada às demandas por agilidade, centralização e confiabilidade da informação orientou a inclusão de questões que abordam variações de consumo, risco de desabastecimento e possibilidade de vencimento sem previsão de uso. Essas informações, quando acessadas de maneira centralizada, contribuem para decisões mais oportunas e fundamentadas.

Considerando que a ferramenta poderá ser utilizada tanto por profissionais do nível gerencial quanto do nível operacional assistencial e que sua aplicação apresenta potencial de utilidade para ambos, foram incluídas perguntas voltadas a temas estratégicos, tais como a estimativa do impacto financeiro de perdas, o monitoramento de medicamentos e materiais de alto custo e o apoio ao planejamento do ressurgimento.

O conjunto de perguntas elaborado para o protótipo está apresentado no Quadro 2.

Quadro 2 – Conjunto inicial de perguntas para o protótipo do *chatbot*

Pergunta	Intenção	Tipo de dado necessário
Quais itens estão com estoque crítico?	Consulta Estoque Crítico	Dados de estoque e parâmetros de criticidade (estoque < 15 dias)
Liste os medicamentos/materiais que vencem nos próximos 90 dias.	Validade Medicamentos	Estoque e datas de validade na nota fiscal
Qual o histórico de consumo do medicamento/material X nos últimos 6 meses?	Histórico Consumo	Dados de consumo por período
Quantas unidades do item X estão disponíveis em estoque agora?	Consulta Estoque Atual	Dados de estoque atual
Há medicamentos/materiais em estoque com consumo abaixo da média histórica?	Identificação de Consumo Baixo	Movimento de estoque e dados de estoque atual
Quais itens tiveram aumento de consumo nos últimos 3 meses?	Tendência Consumo	Histórico de consumo
Quais os 10 medicamentos/materiais mais consumidos no último mês.	Top Medicamentos	Histórico de consumo
Existe algum medicamento/material em falta no estoque hoje?	Medicamento em Falta	Estoque atual

Quais itens estão próximos de vencer e precisam de redistribuição?	Redistribuição Estoque	Estoque e datas de validade na nota fiscal
Qual setor tem maior consumo de um determinado item?	Consumo	Movimento de estoque e local (estoque) da baixa do item
Qual o lote em estoque dos medicamentos/materiais que vencem nos próximos 30 dias?	Validade Medicamentos	Estoque atual, lote e datas de validade na nota fiscal
Qual o impacto financeiro estimado de perdas e vencimentos nos últimos 12 meses?	Avaliação de Custo de Perdas	Movimentações de estoque com perdas + Dados de custo médio unitário do item
Quais itens representam risco de desabastecimento com base no consumo do último trimestre?	Risco Desabastecimento	Estoque atual + histórico de consumo
Quais medicamentos/materiais de alto custo estão com estoque acima de 60 dias de consumo?	Estoque Alto Custo	Dados de consumo, estoque, parâmetros de alto custo (> que R\$300,00/unidade)
Quais medicamentos/materiais têm risco de vencer nos próximos 90 dias sem previsão de consumo?	Prevenção de Perdas por Validade	Datas de validade na nota fiscal + estoque atual + histórico de consumo
Qual a quantidade mínima em estoque do item para o próximo mês de acordo com o consumo médio?	Previsão Consumo	Cálculo do Consumo Médio Mensal (últimos 3 meses) + parâmetro de segurança (+25%)
Quais itens devem ser solicitados no ponto de ressuprimento do próximo mês de acordo com a média de consumo dos últimos 3 meses?	Pedido de Ressuprimento	Estoque atual - Cálculo do Consumo Médio Mensal (últimos 3 meses) + parâmetro de segurança (+25%)
Quais itens registraram perdas no último mês e qual foi a quantidade perdida?	Controle de Perdas	Movimentações de estoque com registro de perdas e quantidades
Quanto solicitar do item no próximo ponto de ressuprimento de acordo com o consumo médio mensal associado ao % de segurança?	Pedido de Ressuprimento	Estoque atual - Cálculo do Consumo Médio Mensal (últimos 3 meses) + parâmetro de segurança (+25%)
Quais medicamentos/materiais tiveram variações de estoque significativas na última semana e qual foi a causa (consumo, perdas, redistribuição ou ajustes de inventário)?	Monitoramento estratégico de estoque e prevenção de desabastecimento	Estoque atual + Movimentações de estoque (Registros de consumo, perdas, redistribuição e ajustes de inventários)

Fonte: Elaborado pela autora (2026).

6.3 Desenvolvimento do protótipo

O desenvolvimento do protótipo foi conduzido utilizando o modelo Qwen3-32B-AWQ, com implementação em linguagem Python, sendo avaliado a partir do conjunto estruturado de 20 perguntas representativas de demandas reais da gestão de estoque

e apoio à tomada de decisão dos processos relacionados às atividades desempenhadas pelo farmacêutico no ambiente hospitalar.

6.3.1 Resultados globais do protótipo e análise geral

De forma global, o protótipo apresentou taxa de sucesso de execução de 85,0%, com 17 consultas executadas com sucesso e 3 consultas resultando em erro de execução. Esse resultado confirma que, do ponto de vista sintático e estrutural, o modelo foi capaz de gerar consultas SQL válidas na maior parte dos cenários avaliados, o que é coerente com o objetivo desta fase do projeto, voltada à validação funcional do protótipo (Pedroso; Pereira; Pereira, 2025; Luo *et al.*, 2025).

As métricas agregadas de desempenho apontaram precisão média de 38,45%, e *recall* médio de 55,59%. Esses valores revelam uma assimetria importante entre a capacidade do modelo de recuperar informações relevantes (*recall*) e sua capacidade de restringir os resultados ao conjunto estritamente esperado (precisão). Tal comportamento já era esperado à luz da complexidade do esquema do banco de dados hospitalar caracterizado por múltiplas tabelas, alto volume de registros e regras de negócio implícitas. Adicionalmente, esses resultados devem ser analisados considerando o curto período de execução efetiva desta fase do projeto, bem como o caráter inovador e de fronteira da proposta, que envolve a aplicação de modelos de linguagem voltados à geração automática de consultas SQL em um contexto de gestão de um estoque hospitalar real, ainda pouco explorado na literatura, reforçando a necessidade de continuidade e aprofundamento do desenvolvimento do protótipo.

Observa-se que a elevada taxa de sucesso de execução não implica, necessariamente, elevada correspondência entre os resultados retornados pelo modelo e o *ground truth*, especialmente em consultas que exigem interpretações mais complexas do contexto de aplicação.

6.3.2 Análise por grupo de perguntas

6.3.2.1 Consultas de estoque e disponibilidade imediata

As consultas relacionadas à verificação direta do estoque atual apresentaram desempenho consistente. A pergunta referente à quantidade disponível de um item específico em estoque (Pergunta 4) obteve correspondência exata entre o resultado do modelo e o resultado de referência, com precisão, *recall* e acurácia de 100%

(Figura 5). Esse resultado confirma que o protótipo é capaz de lidar adequadamente com consultas diretas, baseadas em agregações simples e filtros objetivos.

Figura 5 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 4)

Pergunta 4: Quantas unidades do item X estão disponíveis em estoque agora?

Intenção: Consulta Estoque Atual
Dados necessários: Dados de estoque atual

Parâmetros:
 X1 : 37

Métricas

Tipo: quantidade	Status GT: sucesso	Status Modelo: sucesso	Linhas Esperadas: 1	Linhas Retornadas: 1	Precisão: 100.00%
Recall: 100.00%	Accuracy: 100.00%	F1: 100.00%	Match: ☒ Acertou		

SQL Ground Truth

```

-----
-- Quantas unidades do item X estão disponíveis em estoque agora?
-- 4_m.sql

SELECT SUM(qt_estoque_disp) soma_estoque
FROM estoque_mat_med
WHERE cd_material = 37 ;

```

SQL Modelo

```

SELECT e.cd_material,
       SUM(e.qt_estoque_disp) AS total_disponivel
FROM estoque_mat_med e
WHERE e.cd_material = 37
GROUP BY e.cd_material;

```

Preview Ground Truth (10 primeiras linhas)

soma_estoque
754710

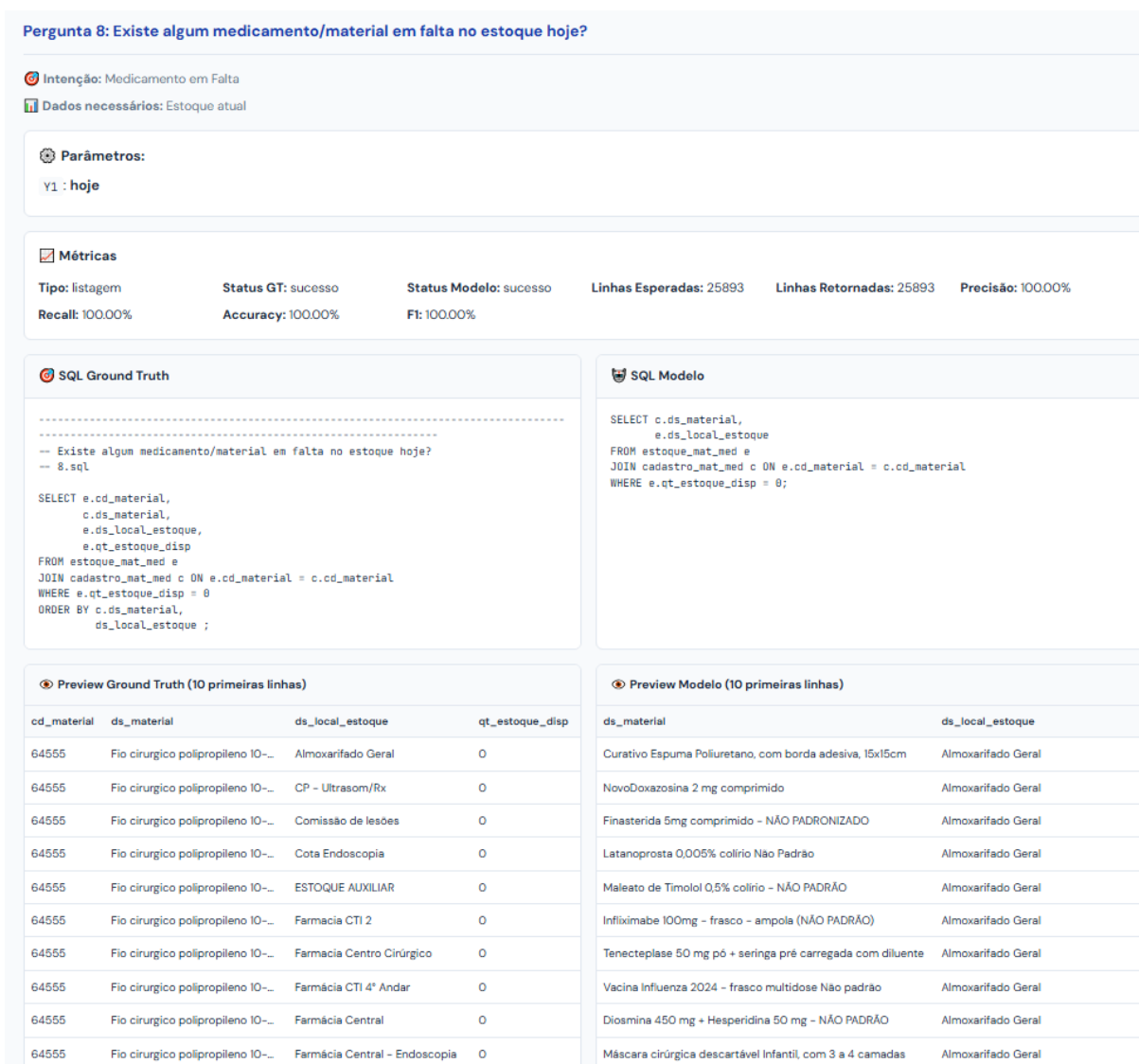
Preview Modelo (10 primeiras linhas)

cd_material	total_disponivel
370	754710

Fonte: Elaborado pela autora (2026).

De modo semelhante, a identificação de medicamentos ou materiais em falta no estoque (Pergunta 8) apresentou recuperação integral das linhas esperadas, indicando alinhamento completo com o *ground truth* (Figura 6). Esses achados reforçam a adequação do protótipo para apoiar atividades operacionais rotineiras da farmácia hospitalar.

Figura 6 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 8)

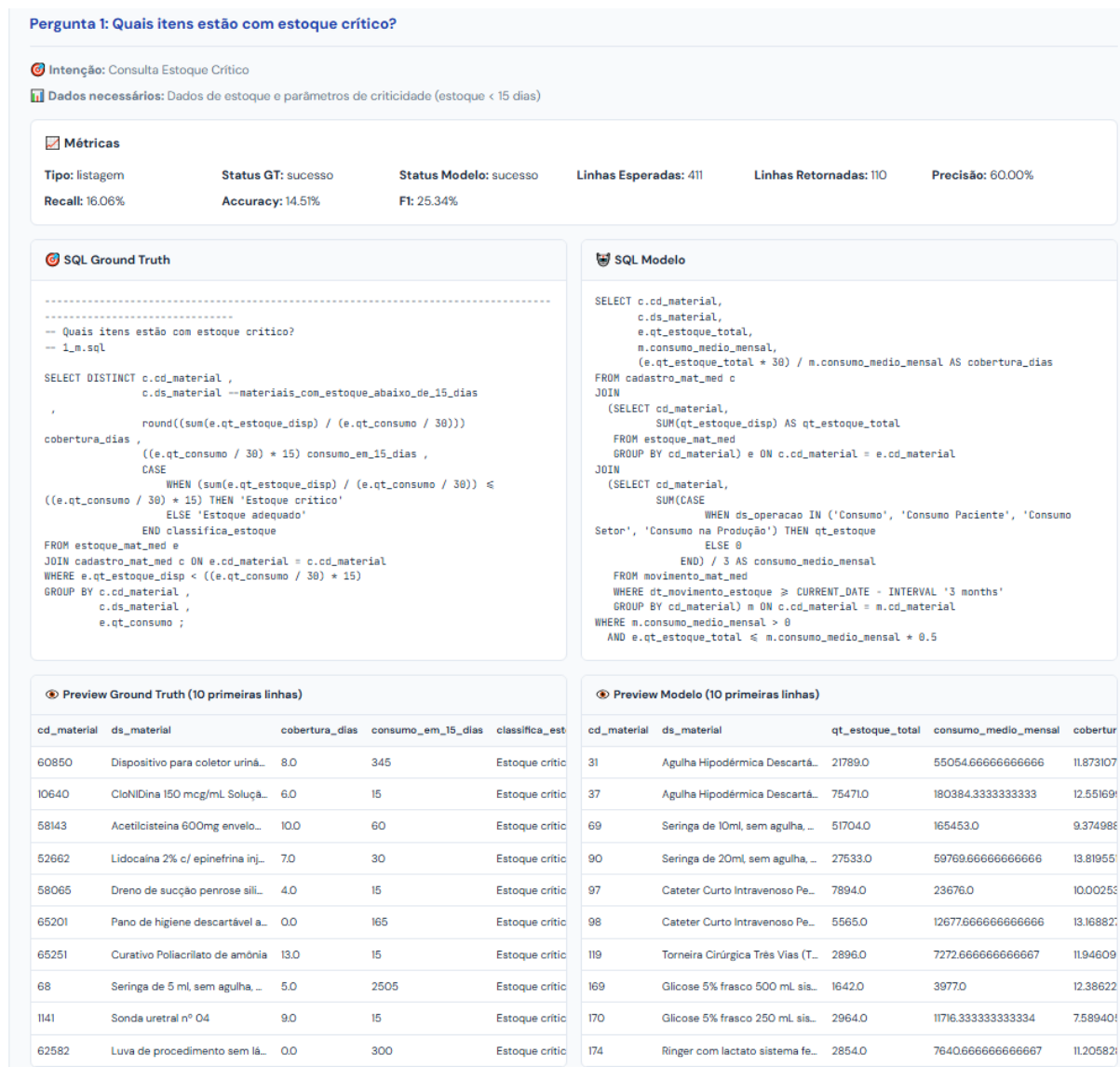


Fonte: Elaborado pela autora (2026).

Em contraste, a consulta destinada à identificação de itens em estoque crítico (Pergunta 1) apresentou desempenho inferior, com precisão de 60,00% e *recall* de 16,06% (Figura 7). A análise do SQL gerado evidencia divergências nos critérios utilizados para definição de criticidade, em relação aos parâmetros explicitados no *ground truth* (cobertura inferior a 15 dias). Esse resultado evidencia a dificuldade do modelo em internalizar regras de negócio implícitas, achado que se mostra consistente com resultados reportados na literatura recente sobre aplicações de modelos de linguagem para geração de consultas SQL em cenários especializados, nos quais limitações semelhantes são observadas quando critérios operacionais não

estão explicitamente formalizados no esquema do banco de dados (Pedroso; Pereira; Pereira, 2025).

Figura 7 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 1)

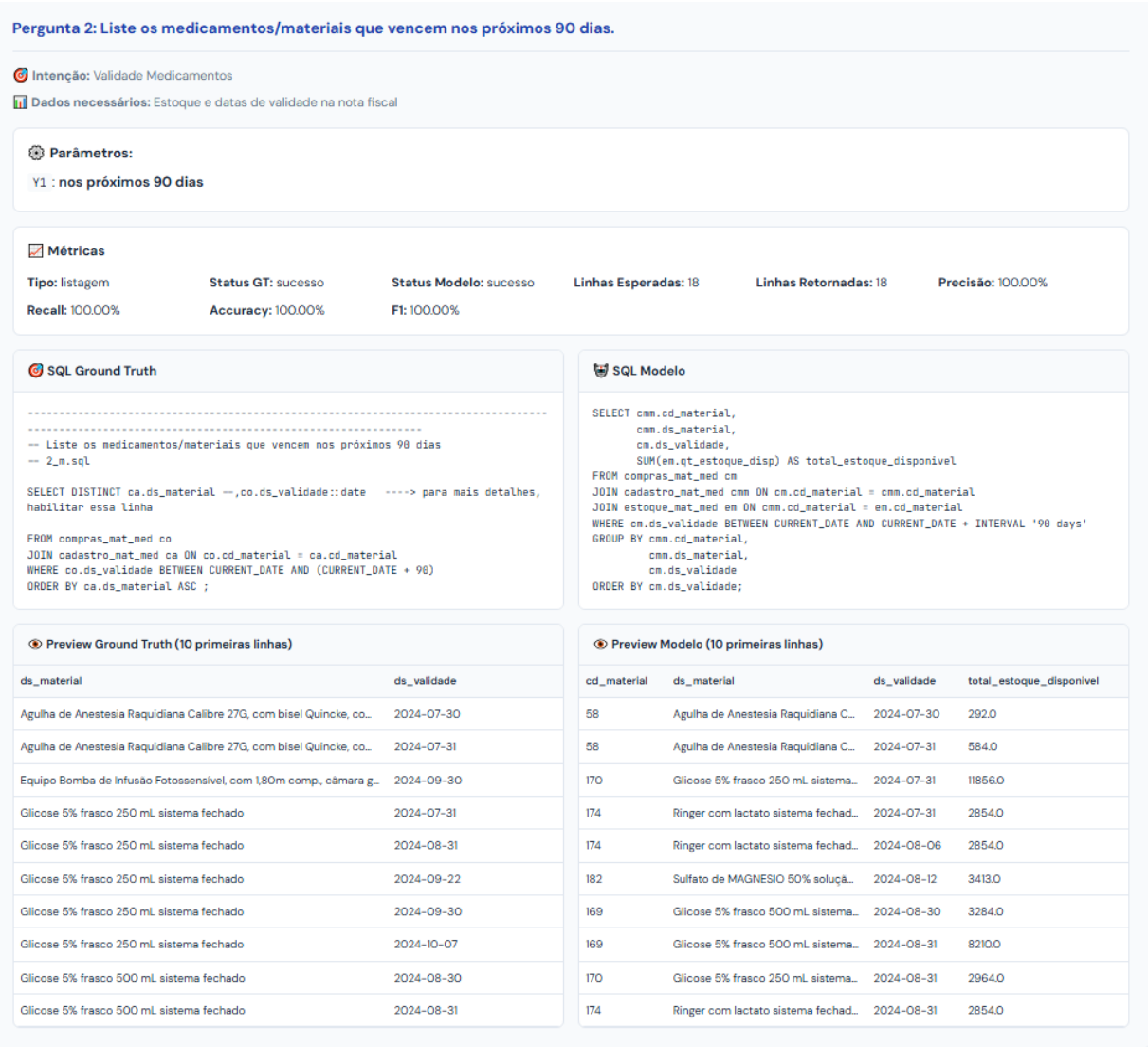


Fonte: Elaborado pela autora (2026).

6.3.2.2 Consultas relacionadas à validade de medicamentos e materiais

As consultas relacionadas à validade apresentaram comportamento heterogêneo. A listagem de medicamentos e materiais com vencimento nos próximos 90 dias foi executada corretamente, com precisão e *recall* de 100% (Figura 8), demonstrando que o modelo conseguiu interpretar adequadamente restrições temporais explícitas quando estas estavam claramente definidas.

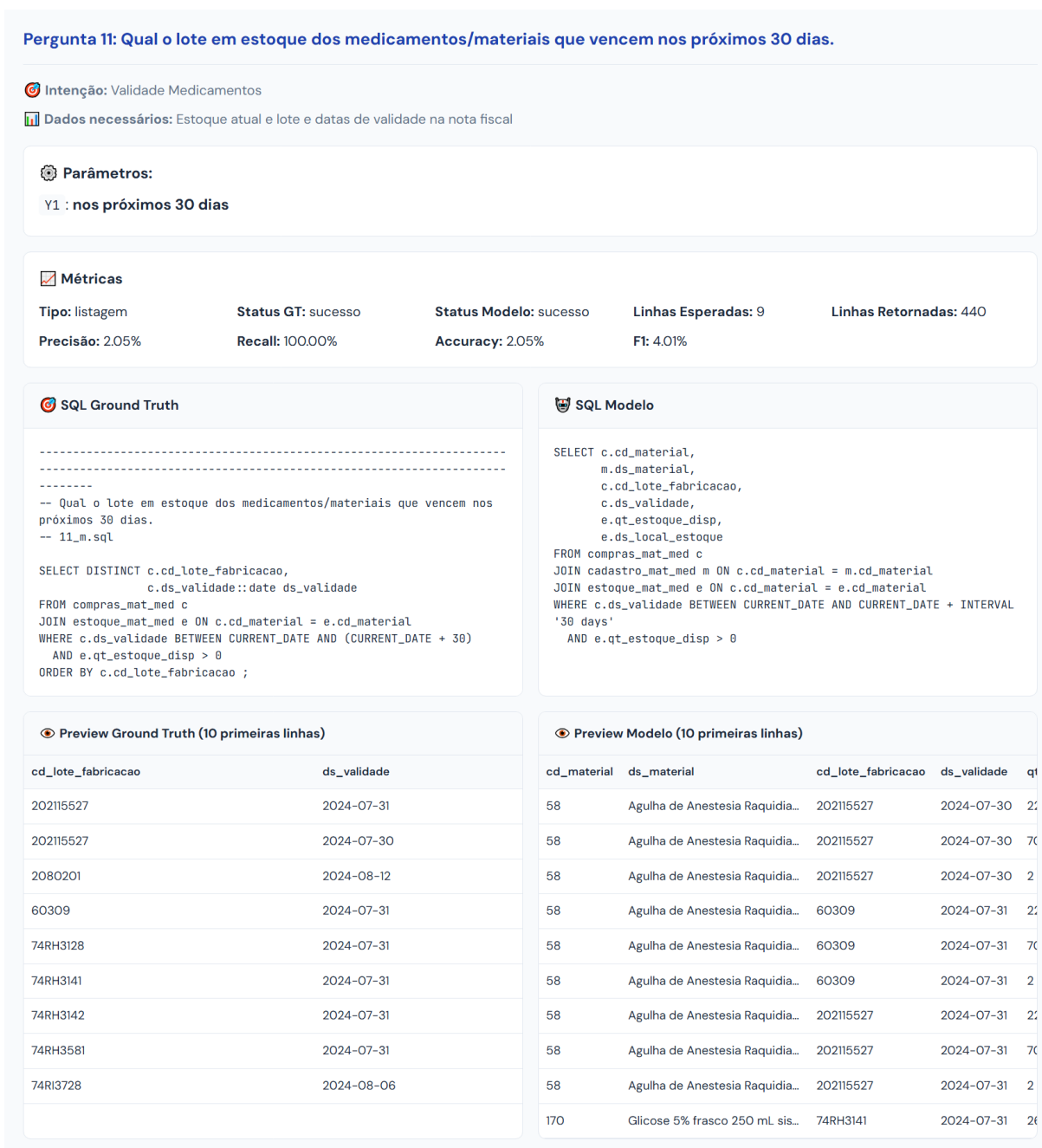
Figura 8 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 2)



Fonte: Elaborado pela autora (2026).

Por outro lado, consultas que exigiam maior detalhamento de informação, como a identificação de lotes específicos de medicamentos com vencimento próximo ou a avaliação da necessidade de redistribuição de itens, apresentaram queda acentuada de precisão. Na consulta sobre lotes com vencimento nos próximos 30 dias (Pergunta 11), embora o *recall* tenha sido de 100%, a precisão foi de apenas 2,05%, em razão do retorno de um número excessivo de registros (Figura 9).

Figura 9 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 11)



Fonte: Elaborado pela autora (2026).

Esse padrão sugere que, embora o modelo consiga recuperar todas as informações relevantes, ele apresenta limitações na aplicação simultânea de múltiplos filtros e junções, especialmente quando envolvem atributos como lote, validade e localização de estoque.

6.3.2.3 Consultas de consumo e histórico temporal

As consultas relacionadas ao histórico de consumo apresentaram alguns dos melhores desempenhos observados nesta fase. A pergunta referente ao histórico de consumo (Pergunta 3) de um item específico nos últimos seis meses obteve *F1-score* de 99,48% (Figura 10), indicando elevada correspondência entre os resultados retornados e o *ground truth*. Resultado igualmente satisfatório foi observado na identificação dos dez medicamentos ou materiais mais consumidos no último mês, com métricas máximas em todos os indicadores avaliados.

Figura 10 – Resultado da consulta SQL gerada pelo protótipo (Pergunta 3)

Pergunta 3: Qual o histórico de consumo do medicamento/material X nos últimos 6 meses?

Intenção: Histórico Consumo

Dados necessários: Dados de consumo por período

Parâmetros:

X1 : 37

Y1 : nos últimos 6 meses

Métricas

Tipo: listagem	Status GT: sucesso	Status Modelo: sucesso	Linhas Esperadas: 384651	Linhas Retornadas: 388641	Precisão: 98.97%
Recall: 100.00%	Accuracy: 98.97%	F1: 99.48%			

SQL Ground Truth

```

-----
-- Qual o histórico de consumo do medicamento/material X nos últimos 6 meses?
-- 3_m.sql

SELECT m.dt_movimento_estoque::date AS DATA ,
       m.dt_movimento_estoque::TIME AS hora ,
       c.ds_material ,
       m.qt_estoque ,
       m.ds_operacao ,
       m.ds_local_estoque
FROM movimento_mat_med m
JOIN cadastro_mat_med c ON m.cd_material = c.cd_material
WHERE m.ds_operacao LIKE '%Consumo%'
AND m.dt_movimento_estoque BETWEEN (CURRENT_DATE - 180) AND CURRENT_DATE
AND c.cd_material = 37
ORDER BY m.dt_movimento_estoque DESC ;

```

SQL Modelo

```

SELECT m.cd_material,
       c.ds_material,
       m.dt_movimento_estoque,
       m.qt_estoque,
       m.ds_local_estoque
FROM movimento_mat_med m
JOIN cadastro_mat_med c ON m.cd_material = c.cd_material
WHERE m.cd_material = 37
AND m.ds_operacao LIKE 'Consumo%'
AND m.dt_movimento_estoque BETWEEN (CURRENT_DATE - INTERVAL '6 months') AND
CURRENT_DATE
ORDER BY m.dt_movimento_estoque DESC;

```

Preview Ground Truth (10 primeiras linhas)

cd_material	dt_movimento_estoque	qt_estoque	ds_operacao	ds_local_estoque
37	2024-07-18 23:58:02	1	Consumo Paciente	Dispensário I - CTI 4
37	2024-07-18 23:54:18	1	Consumo Paciente	Dispensário II - CTI 4
37	2024-07-18 23:52:47	1	Consumo Paciente	Dispensário 7º Sul
37	2024-07-18 23:52:47	1	Consumo Paciente	Dispensário 7º Sul
37	2024-07-18 23:48:18	1	Consumo Paciente	Dispensário 8º Sul
37	2024-07-18 23:42:19	1	Consumo Paciente	Dispensário I - CTI 2
37	2024-07-18 23:42:19	1	Consumo Paciente	Dispensário I - CTI 2
37	2024-07-18 23:40:49	1	Consumo Paciente	Dispensário I - CTI 2
37	2024-07-18 23:37:49	1	Consumo Paciente	Dispensário 6º Sul
37	2024-07-18 23:35:35	1	Consumo Paciente	Dispensário I - CTI 2

Preview Modelo (10 primeiras linhas)

cd_material	ds_material	dt_movimento_estoque	qt_estoque	ds_local_estoque
37	Agulha Hipodérmica Descartá...	2024-07-18 23:58:02	1	Dispensário I - CTI 4
37	Agulha Hipodérmica Descartá...	2024-07-18 23:54:18	1	Dispensário II - CTI 4
37	Agulha Hipodérmica Descartá...	2024-07-18 23:52:47	1	Dispensário 7º Sul
37	Agulha Hipodérmica Descartá...	2024-07-18 23:52:47	1	Dispensário 7º Sul
37	Agulha Hipodérmica Descartá...	2024-07-18 23:48:18	1	Dispensário 8º Sul
37	Agulha Hipodérmica Descartá...	2024-07-18 23:42:19	1	Dispensário I - CTI 2
37	Agulha Hipodérmica Descartá...	2024-07-18 23:42:19	1	Dispensário I - CTI 2
37	Agulha Hipodérmica Descartá...	2024-07-18 23:40:49	1	Dispensário I - CTI 2
37	Agulha Hipodérmica Descartá...	2024-07-18 23:37:49	1	Dispensário 6º Sul
37	Agulha Hipodérmica Descartá...	2024-07-18 23:35:35	1	Dispensário I - CTI 2

Fonte: Elaborado pela autora (2026).

Fonte: Elaborado pela autora (2026).

6.3.2.4 Consultas de apoio à decisão e gestão estratégica

As consultas associadas ao apoio à decisão gerencial apresentaram os menores desempenhos globais. Embora muitas dessas consultas tenham sido executadas sem erro técnico, os resultados retornados divergiram substancialmente do *ground truth*.

Na avaliação do impacto financeiro de perdas e vencimentos nos últimos 12 meses, por exemplo, o modelo retornou um valor único, enquanto o resultado de referência previa a discriminação por tipo de operação. Situação semelhante foi observada nas consultas relacionadas à estimativa de quantidades para ressuprimento e definição de estoque mínimo, cujas métricas de precisão e *recall* foram iguais a zero.

Esses achados evidenciam que consultas que incorporam múltiplas regras de negócio, parâmetros financeiros e projeções temporais representam um desafio significativo para o protótipo nesta fase de desenvolvimento.

6.3.2.5 Consultas com erro de execução

Três consultas resultaram em erro de execução, todas associadas a perguntas que combinavam critérios complexos, como custo unitário elevado, consumo médio e janelas temporais prolongadas. Esses erros indicam limitações do modelo na geração de consultas SQL válidas quando o nível de complexidade semântica aumenta, reforçando a necessidade de refinamento adicional do protótipo.

6.3.3 Síntese integrada dos resultados e discussão

De forma integrada, conforme apresentado na Figura 12, os resultados desta fase demonstram que o protótipo apresenta desempenho superior em consultas operacionais diretas, baseadas em agregações simples e filtros explícitos, especialmente aquelas relacionadas à verificação de estoque e ao histórico de consumo. Esse comportamento está alinhado com avaliações prévias de modelos de linguagem aplicados à tarefa de conversão de linguagem natural para consultas SQL, que indicam maior robustez em cenários de menor complexidade semântica e com regras de negócio claramente definidas (Pedroso; Pereira; Pereira, 2025).

No desenvolvimento tecnológico, o protótipo apresentou taxa global de sucesso de execução de 85%, evidenciando que, do ponto de vista sintático e estrutural, o modelo foi capaz de gerar consultas SQL válidas na maioria dos cenários testados.

Entretanto, as métricas de precisão média (38,45%), *recall* (55,59%) e *F1-score* (36,21%) revelaram assimetria entre a capacidade de recuperar informações relevantes e a capacidade de restringir adequadamente os resultados ao conjunto esperado.

A análise por grupo de perguntas evidenciou desempenho superior do protótipo em consultas operacionais diretas, baseadas em agregações simples e filtros explícitos, como verificação de estoque, identificação de itens em falta, consultas de validade com restrições temporais claras e histórico de consumo. Nesses cenários, foram observados casos de correspondência integral com o resultado de referência, reforçando a adequação do protótipo para apoiar atividades operacionais rotineiras da farmácia hospitalar.

Em contrapartida, as consultas relacionadas ao apoio à decisão gerencial e às regras de negócio implícitas, como definição de estoque crítico, cálculo de ponto de ressuprimento, análises financeiras e projeções temporais, apresentaram desempenho inferior, com baixos valores de precisão e *F1-score*, além de ocorrências de erro de execução. Esses resultados indicam limitações do modelo em internalizar conhecimento normativo e regras que não estão explicitamente formalizadas no esquema do banco de dados, o que é consistente com achados da literatura recente sobre aplicações de LLMs em domínios especializados (Pedroso; Pereira; Pereira, 2025; Luo *et al.*, 2025; Raimundo; Dias; Guerra, 2015; Barbosa, 2015; Raza *et al.*, 2022).

Por outro lado, os problemas encontrados em consultas que dependem de regras que não estão claramente descritas no sistema, como estoque crítico, ponto de ressuprimento e análises comparativas de consumo, confirmam o que a literatura descreve. Esses estudos indicam que modelos de conversão de linguagem natural para consultas SQL costumam ter dificuldade para lidar com informações de contexto e regras internas que não aparecem de forma explícita na estrutura do banco de dados (Pedroso; Pereira; Pereira, 2025; Luo *et al.*, 2025).

A diferença observada entre a alta taxa de execução bem-sucedida das consultas e os níveis apenas moderados de precisão reforça a distinção entre correção sintática e aderência semântica, já amplamente discutida na literatura sobre avaliação de sistemas que convertem linguagem natural em consultas SQL. Esses resultados sugerem que a geração de consultas SQL tecnicamente válidas não

assegura, por si só, a recuperação de dados que correspondam, de fato, à intenção e às expectativas do usuário final (Pedroso; Pereira; Pereira, 2025; Luo *et al.*, 2025).

Além disso, os resultados precisam ser analisados considerando a complexidade própria do ambiente hospitalar, especialmente no que se refere à gestão farmacêutica. Como apontado pela literatura, trata-se de um contexto marcado pela presença de diversas variáveis operacionais, financeiras e regulatórias, o que torna mais desafiadora a aplicação de modelos de linguagem em cenários reais e altamente especializados (Raimundo; Dias; Guerra, 2015; Barbosa, 2015; Raza *et al.*, 2022).

Figura 12 – Desempenho do protótipo por consulta

Resumo por Pergunta			
ID	PERGUNTA	TIPO	RESULTADO
1	Quais itens estão com estoque crítico?	listagem	✗ Incorreto
2	Liste os medicamentos/materiais que vencem nos próximos 90 dias.	listagem	✓ Correto
3	Qual o histórico de consumo do medicamento/material X nos últimos 6 meses?	listagem	✓ Correto
4	Quantas unidades do item X estão disponíveis em estoque agora?	quantidade	✓ Correto
5	Há medicamentos/materiais em estoque com consumo abaixo da média histórica?	listagem	✗ Incorreto
6	Quais itens tiveram aumento de consumo nos últimos 3 meses?	listagem	✗ Incorreto
7	Quais os 10 medicamentos/materiais mais consumidos no último mês.	listagem	✓ Correto
8	Existe algum medicamento/material em falta no estoque hoje?	listagem	✓ Correto
9	Quais itens estão próximos de vencer e precisam de redistribuição?	listagem	✗ Erro
10	Qual setor tem maior consumo de um determinado item?	quantidade	✗ Incorreto
11	Qual o lote em estoque dos medicamentos/materiais que vencem nos próximos 30 dias...	listagem	✗ Incorreto
12	Qual o impacto financeiro estimado de perdas e vencimentos nos últimos 12 meses?	quantidade	✗ Incorreto
13	Quais itens representam risco de desabastecimento com base no consumo do último ...	listagem	✗ Incorreto
14	Quais medicamentos/materiais de alto custo estão com estoque acima de 60 dias de...	listagem	✗ Erro
15	Quais medicamentos/materiais têm risco de vencer nos próximos 90 dias sem previs...	listagem	✗ Incorreto
16	Qual a quantidade mínima em estoque do item para o próximo mês de acordo com o c...	quantidade	✗ Incorreto
17	Quais itens devem ser solicitados no ponto de ressuprimento do próximo mês de ac...	listagem	✗ Erro
18	Quais itens registraram perdas no último mês e qual foi a quantidade perdida?	listagem	✗ Incorreto
19	Quanto solicitar do item no próximo ponto de ressuprimento de acordo com o consu...	quantidade	✗ Incorreto
20	Quais medicamentos/materiais tiveram variações de estoque significativas na últi...	listagem	✗ Incorreto

Fonte: Elaborado pela autora (2026).

Esses elementos reforçam que os achados da fase de desenvolvimento da ferramenta tecnológica estão em consonância com o estado da arte, evidenciando tanto o potencial do protótipo para consultas operacionais quanto as limitações esperadas em análises gerenciais mais complexas, o que fornece subsídios importantes para as etapas subsequentes de refinamento e validação do sistema.

6.4 Limitações e desafios enfrentados no desenvolvimento do protótipo

O desenvolvimento do protótipo enfrentou desafios significativos, especialmente relacionados à complexidade técnica do projeto, ao financiamento e à interação interdisciplinar

Um desafio relevante decorreu da natureza inovadora e interdisciplinar do projeto, que envolveu áreas de conhecimento bastante distintas: a farmacêutica e a tecnológica. O desenvolvimento do protótipo demandou competências avançadas em programação. Assim, as atividades restritas à área farmacêutica puderam ser iniciadas imediatamente após a emissão do parecer favorável do CEP, enquanto a implementação técnica ficou condicionada à colaboração da equipe de TI, que somente pôde ser constituída após a liberação do financiamento aprovado.

A comunicação das regras de negócio da farmácia para a equipe de TI também apresentou complexidade significativa, em função da distância entre os diferentes domínios de conhecimento: conceitos cotidianos da prática farmacêutica eram pouco familiares à equipe de TI, assim como a programação representava um desafio para a compreensão da pesquisadora. Para contornar essas dificuldades, foram realizadas reuniões semanais de alinhamento, possibilitando a construção colaborativa do protótipo e a mitigação de problemas relacionados à integração entre as regras de negócio farmacêuticas e sua implementação técnica.

Problemas relacionados ao banco de dados também representaram entraves significativos, consistentes com desafios relatados na literatura sobre desenvolvimento de protótipos inovadores em saúde digital, segundo a qual a experimentação e a integração de dados complexos são reconhecidamente obstáculos comuns (Bellei *et al.*, 2025; Pedroso; Pereira; Pereira, 2025).

Embora significativos, esses entraves não decorreram de negligência ou de falhas de planejamento por parte da equipe de pesquisa. Ao contrário, refletem a natureza inovadora e interdisciplinar do projeto, que demandou a adoção de estratégias de mitigação ativas, incluindo reuniões periódicas com a equipe de

tecnologia da informação, documentação detalhada das regras de negócio e das informações extraídas do sistema eletrônico utilizado no hospital, bem como planejamento criterioso do uso dos recursos financeiros. Essas medidas permitiram o desenvolvimento do protótipo dentro das limitações impostas, garantindo a continuidade do projeto e o potencial de aprimoramento em etapas futuras.

6.5 Perspectivas futuras

O desenvolvimento do protótipo representou um avanço na aplicação de PLN para apoiar a gestão farmacêutica em ambiente hospitalar. Apesar das limitações enfrentadas, o projeto apresenta diversas possibilidades de continuidade e expansão, que incluem:

1. Aprimoramento do protótipo com vistas à resolução das perguntas que, na etapa atual do estudo, ainda não apresentaram resultados plenamente satisfatórios.
2. Exploração de consultas adicionais: Ampliar o escopo do protótipo para incluir outras consultas relacionadas à gestão de estoque. Essa expansão permitirá uma análise mais abrangente e integrada da cadeia de suprimentos farmacêuticos, além de possibilitar a manutenção e atualização contínua do código-fonte no repositório GitHub, garantindo que o protótipo permaneça funcional, documentado e disponível para futuras modificações e melhorias.
3. Treinamento de diferentes modelos de linguagem (LLMs): Avaliar e treinar outros Modelos Amplos de Linguagem, visando melhorar a acurácia das respostas e a capacidade de compreensão de consultas complexas.
4. Validação e avaliação em ambiente real: Realizar testes de usabilidade e avaliação de desempenho do protótipo com profissionais da farmácia hospitalar, ajustando funcionalidades e interface de acordo com o *feedback* recebido.
5. Integração com sistemas hospitalares: Avançar na integração do protótipo com sistemas eletrônicos hospitalares existentes, aprimorando a interoperabilidade e permitindo que o *chatbot* acesse dados em tempo real, aumentando a eficiência e utilidade do sistema.

6. Expansão para suporte a decisões clínicas e administrativas: Explorar a utilização do protótipo em consultas que vão além da gestão de estoque farmacêutico, mas também no acesso facilitado à informação relacionado a outros setores em um hospital.

O protótipo, portanto, apresenta-se como uma ferramenta potencialmente estratégica para apoiar a tomada de decisão, otimizar o uso de recursos e qualificar os processos da gestão farmacêutica hospitalar, na medida em que favorece o acesso estruturado, ágil e contextualizado às informações operacionais (Raimundo; Dias; Guerra, 2015; Barbosa, 2015; Nascimento Neto *et al.*, 2020; Farias Leal *et al.*, 2022; Pedroso; Pereira; Pereira, 2025; Teixeira; Andrade, 2022; Raza *et al.*, 2022; Luo *et al.*, 2025). Tal perspectiva encontra aderência às diretrizes de eficiência, segurança, rastreabilidade e racionalidade preconizadas pelas políticas públicas de saúde e pela literatura sobre gestão hospitalar e inovação em saúde. Ressalta-se, contudo, que a consolidação desse potencial depende da evolução do protótipo, da integração com sistemas institucionais, do fortalecimento da governança de dados e da validação em ambiente real, de modo a assegurar confiabilidade, sustentabilidade e alinhamento às práticas organizacionais (Brasil, 2010; Teixeira; Andrade, 2022).

7 CONCLUSÃO

O presente estudo teve como propósito explorar a aplicação de Modelos Amplos de Linguagem, ou LLMs, no contexto da gestão farmacêutica hospitalar, com foco no desenvolvimento de um protótipo capaz de apoiar a recuperação de informações e a tomada de decisão do profissional farmacêutico. Ao longo do trabalho, ficou evidente que a complexidade dos processos envolvidos na gestão de estoques, aliada à diversidade de demandas do cotidiano hospitalar, impõe desafios significativos à organização e ao acesso eficiente às informações.

A realização do estudo em duas fases complementares mostrou-se fundamental para a construção de uma solução alinhada à prática profissional. A etapa de pesquisa e planejamento, conduzida com profissionais que vivenciam rotineiramente os processos da gestão farmacêutica, permitiu identificar dificuldades recorrentes no acesso às informações, fragmentação de sistemas e dependência de processos manuais, assim como descrito na literatura sobre os desafios estruturais da gestão da assistência farmacêutica hospitalar (Brasil, 2010; Teixeira; Andrade, 2022; Farias Leal *et al.*, 2022). Esses resultados permitiram definir, de forma contextualizada, o conjunto de perguntas que orientou o desenvolvimento do protótipo.

Os resultados obtidos indicam que a utilização de um modelo de linguagem, associada a uma estratégia de recuperação de informações estruturadas, apresenta potencial para contribuir com a organização do conhecimento e com o apoio à tomada de decisão no âmbito da gestão farmacêutica hospitalar. Ainda que algumas consultas não tenham alcançado resultados plenamente satisfatórios na etapa atual, o protótipo demonstrou viabilidade conceitual e funcional, evidenciando caminhos promissores para seu aprimoramento futuro.

Mais do que apresentar uma solução tecnológica acabada, este estudo contribuiu para refletir sobre o papel do farmacêutico frente às inovações em inteligência artificial, destacando a necessidade de integração entre conhecimento técnico, prática profissional e tecnologias emergentes. Nesse sentido, o trabalho reafirma que o valor dessas ferramentas não reside apenas em sua sofisticação computacional, mas sobretudo em sua capacidade de apoiar o cuidado, otimizar processos e fortalecer a atuação do profissional de saúde no contexto hospitalar.

Conclui-se que o protótipo desenvolvido responde de forma consistente à pergunta norteadora da pesquisa, ao demonstrar que é tecnicamente possível estruturar um *chatbot* baseado em PLN para organizar e disponibilizar informações da gestão de insumos farmacêuticos. Os resultados confirmam o potencial da solução para apoiar rotinas operacionais e evidenciam, simultaneamente, os desafios associados à aplicação de modelos de linguagem em cenários decisórios complexos. O trabalho contribui para o avanço do campo das tecnologias gerenciais em saúde, ao integrar fundamentos teóricos, desenvolvimento tecnológico e avaliação empírica, oferecendo subsídios tanto para pesquisas futuras quanto para iniciativas institucionais de inovação na gestão farmacêutica hospitalar.

REFERÊNCIAS

- AGUIAR, Rafael A. T. *As áreas da inteligência artificial*. 2024a. Trabalho não publicado. Disponível em: <https://raphaelaguilar.notion.site/Deep-Learning-e-Vis-o-Computacional-64091502ebe74e13911c3d0e5fc8d874>. Acesso em: 31 maio 2024.
- AGUIAR, Raphael A. T. *Deep learning e visão computacional*. 2024b. Trabalho não publicado. Disponível em: <https://raphaelaguilar.notion.site/Deep-Learning-e-Vis-o-Computacional-64091502ebe74e13911c3d0e5fc8d874>. Acesso em: 31 maio 2024.
- ALMEIDA, Celia. Parcerias público-privadas (PPP) no setor saúde: processos globais e dinâmicas nacionais. *Cadernos de Saúde Pública*, v. 33, n. Supl. 2, e00197316, 2017. Doi: <https://doi.org/10.1590/0102-311X00197316>. Acesso em: 14 mar. 2026.
- AMORIM, Maurício J. V.; BARONE, Dante; MANSUR, André Uebe. Técnicas de aprendizado de máquina aplicadas na previsão de evasão acadêmica. *In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE)*, 19, 2008. *Anais [...]*. [S. l], 2008. p. 666–674.
- BARBOSA, Kátia Simone da Silva. Gerenciamento de farmácia hospitalar: otimização da qualidade, produtividade e recursos financeiros. *Revista Saúde e Desenvolvimento*, v. 7, n. 4, p. 1–10, jan./dez. 2015.
- BARDIN, Laurence. *Análise de conteúdo*. Lisboa: Edições 70, 1977.
- BELLEI, Ericles Andrei; MORETTO, Cleide Fátima; FREITAS, Carla Maria Dal Sasso; DE MARCHI, Ana Carolina Bertoletti. A Survey on Digital Solutions for Health Services Management: Features and Use Cases from Brazilian National Literature. *Healthcare*, Basel, v. 13, n. 18, 18 set. 2025. Doi: 10.3390/healthcare13182348. Acesso em: 3 jan. 2026.
- BENNING, Leo; PEINTNER, Andreas; PEINTNER, Lukas. Advances in and the Applicability of Machine Learning-Based Screening and Early Detection Approaches for Cancer: A Primer. *Cancers*, v. 14, n. 3, p. 623, 2022. Doi: <https://doi.org/10.3390/cancers14030623>. Acesso em: 31 maio 2024.
- BJARNASON, Elizabeth; LANG, Franz; MJÖBERG, Alexander. An Empirically Based Model of Software Prototyping: A Mapping Study and a Multi-Case Study. *Empirical Software Engineering*, v. 28, n. 5, art. 115, 2023. Doi: <https://doi.org/10.1007/s10664-023-10331-w>. Acesso em: 14 mar. 2026
- BRASIL. Conselho Nacional de Saúde. *Resolução nº 338, de 6 de maio de 2004*. Aprova a Política Nacional de Assistência Farmacêutica e dá outras providências. Brasília, DF: CNS 2004a.
- BRASIL. Conselho Nacional de Secretários de Saúde. *Assistência farmacêutica no SUS*. Brasília: CONASS, 2007. (Coleção Para entender a gestão do SUS, v. 7).
- BRASIL. Ministério da Gestão e da Inovação em Serviços Públicos. Guia prático de prompt e pesquisa com IA para servidores públicos. *In: GOV.BR [Site]*, 2025.

Disponível em: <https://www.gov.br/governodigital/pt-br/infraestrutura-nacional-de-dados/inteligencia-artificial-1/publicacoes/guia-pratico-de-prompt-e-pesquisa-com-ia-para-servidores-publicos>. Acesso em: 17 jan. 2026.

BRASIL. Ministério da Saúde. Conselho Nacional de Saúde. *Resolução nº 466, de 12 de dezembro de 2012*. Aprova novas diretrizes e normas regulamentadoras de pesquisas envolvendo seres humanos; revoga as Res. nº 196/96), Res. nº 303/00, Res. nº 404/08. Brasília: CNS/MS, 2012.

BRASIL. Ministério da Saúde. *Portaria nº 4.283, de 30 de dezembro de 2010*. Aprova as diretrizes e estratégias para organização, fortalecimento e aprimoramento das ações e serviços de farmácia no âmbito dos hospitais. Brasília, DF: Ministério da Saúde, 2010.

BRASIL. Lei nº 11.079, de 30 de dezembro de 2004. Institui normas gerais para licitação e contratação de parceria público-privada no âmbito da administração pública. *Diário Oficial da União*, Brasília, DF, seção 1, ano 141, n. 252, p. 6–8, 31 dez. 2004b.

BRUNS, Suelma de Fátima; LUIZA, Vera Lucia; OLIVEIRA, Egléubia Andrade de. Gestão da assistência farmacêutica em municípios do estado da Paraíba (PB): olhando a aplicação de recursos públicos. *Revista de Administração Pública*, v. 48, n. 3, p. 745–765, 2014. Doi: <https://doi.org/10.1590/0034-76121502>. Acesso em: 31 maio 2024.

BZHALAVA, Levan *et al.* Mapping the wave of industry digitalization by co-word analysis: an exploration of four disruptive industries. *International Journal of Innovation and Technology Management*, v. 19, n. 2, 2022. Doi: <https://doi.org/10.1142/S0219877022500018>. Acesso em: 31 maio 2024.

CASELI, Helena; FREITAS, Cláudia; VIOLA, Roberta. Processamento de linguagem natural. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. *Tópicos em gerenciamento de dados e informações*: miniucursos do SBBB 2022. Organizadores: Ticiane L. Coelho da Silva, Eduardo Ogasawara, Damires Souza, Sérgio Lifschitz. Porto Alegre: Sociedade Brasileira de Computação, 2022. p. 1-25. Doi: <https://doi.org/10.5753/sbc.10309.7.1>. Acesso em: 31 maio 2024.

CONSELHO FEDERAL DE FARMÁCIA (CFF). *Resolução nº 568, de 6 de dezembro de 2012*. Ementa: Dá nova redação aos artigos 1º ao 6º da Resolução/CFF nº 492 de 26 de novembro de 2008, que regulamenta o exercício profissional nos serviços de atendimento pré-hospitalar, na farmácia hospitalar e em outros serviços de saúde, de natureza pública ou privada. Brasília, DF: CFF, 2012.

CONSELHO REGIONAL DE FARMÁCIA DO ESTADO DE SÃO PAULO. *Farmácia hospitalar*. 4. ed. São Paulo: CRF-SP, 2019.

COPELAND, Brian Jack. Artificial intelligence. In: ENCYCLOPEDIA Britannica [site], 30 maio 2024. Disponível em: <https://www.britannica.com/technology/artificial-intelligence>. Acesso em: 31 maio 2024.

DATABRICKS. Geração aumentada de recuperação (RAG). In: DATABRICKS [site] Glossário Databricks, [2025]. Disponível em: <https://www.databricks.com/br/glossary/retrieval-augmented-generation-rag>. Acesso em: 12 jan. 2026.

DATE, Christopher J. *Introdução a sistemas de bancos de dados*. 8. ed. Rio de Janeiro: Elsevier, 2004.

DONG, Shi; WANG, Ping; ABBAS, Khushnood. A Survey on Deep Learning and its Applications. *Computer Science Review*, v. 40, issue c, 2021. Doi: <https://doi.org/10.1016/j.cosrev.2021.100379>. Acesso em: 6 jun. 2024.

DÖRR, Konstantin Nicholas. Mapping the Field of Algorithmic Journalism. *Digital Journalism*, v. 4, n. 6, p. 700–722, 2015. Doi: https://doi.org/10.1080/21670811.2015.1096748?urlappend=%3Futm_source%3DResearchgate.net%26utm_medium%3Darticle. Acesso em: 6 jun. 2024.

FARIAS LEAL, Adriana Amorim *et al.* Atividades clínicas desenvolvidas pelo farmacêutico no contexto da farmácia hospitalar: revisão integrativa. *Acta Farmacêutica Portuguesa*, v. 11, n. 1, p. 98–108, 2022. Disponível em: <https://actafarmacaceuticaportuguesa.com/index.php/afp/article/view/301>. Acesso em: 24 maio 2024.

FENG, Yingchaojie *et al.* Promptmagician: Interactive prompt engineering for text-to-image creation. *IEEE Transactions on Visualization and Computer Graphics*, v. 30, n. 1, p. 295–305, 2024. Doi: <https://doi.org/10.1109/TVCG.2023.3327168>. Acesso em: 3 jan. 2026.

FERNANDES, Fernanda dos Santos; BORDIN, Ronaldo. Desempenho da gestão hospitalar por parcerias público-privadas no Sistema Único de Saúde. *REAd – Revista Eletrônica de Administração*, Porto Alegre, v. 28, n. 3, p. 754–769, set./dez. 2022. Doi: <https://doi.org/10.1590/1413-2311.367.122332> Acesso em: 14 mar. 2026.

FERRI, C.; HERNÁNDEZ-ORALLO, J.; MODROIU, R. An Experimental Comparison of Performance Measures for Classification. *Pattern Recognition Letters*, v. 30, n. 1, p. 27–38, 2009. Doi: <https://doi.org/10.1016/j.patrec.2008.08.010>. Acesso em: 3 jan. 2026.

FLORIDI, Luciano; CHIRIATTI, Massimo. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, v. 30, n. 4, p. 681–694, 2020. Doi: <https://doi.org/10.1007/s11023-020-09548-1>. Acesso em: 17 maio 2024.

HOSHI, Yasuto *et al.* RaLLe: A Framework for Developing and Evaluating Retrieval-Augmented Large Language Models. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING: SYSTEM DEMONSTRATIONS, 2023, Singapura. *Proceedings [...]*. Singapura: Association for Computational Linguistics, 2023. Disponível em: <http://arxiv.org/abs/2308.10633>. Acesso em: 18 jan. 2026.

ILIASHENKO, Oksana; BIKKULOVA, Zilia; DUBGORN, Alissa. Opportunities and Challenges of Artificial Intelligence in Healthcare. *In: INTERNATIONAL SCIENCE CONFERENCESPBWOSCE-2018*, 2018, St. Petersburg. *Proceedings [...]*. St. Petersburg: E3S Web of Conferences, 2019.

JIN, Mingyu *et al.* Health-LLM: Personalized Retrieval-Augmented Disease Prediction System. *arXiv*, 2024. Doi: <https://doi.org/10.48550/arXiv.2402.00746>. Acesso em: 10 jan. 2026.

JOSHI, Diksha. The Role of AI in Pharmacy: A Review. *World Journal of Pharmaceutical Research*, v. 12, n. 12, p. 489–494, 2023. Doi: 10.20959/wjpr202312-28962. Acesso em: 24 maio 2024.

KRANTZ, Tom; JONKER, Alexandra. O que é verdade fundamental? *In: IBM [Site]*, [2025]. Disponível em: <https://www.ibm.com/br-pt/think/topics/ground-truth>. Acesso em: 8 jan. 2026.

LEE, Peter; BUBECK, Sebastien; PETRO, Joseph. Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *The New England Journal of Medicine*, v. 388, n. 13, p. 1233–1239, 2023. Doi: 10.1056/NEJMSr2214184. Acesso em: 24 maio 2024.

LEWIS, Patrick *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS*, 34, 2020, Vancouver. *Proceedings [...]*. Vancouver: NeurIPS, 2020. Disponível em: <https://arxiv.org/abs/2005.11401>. Acesso em: 18 jan. 2026.

LIMA, Rodrigo Fonseca; CÂMARA, Tássio Fontes Moreira. Desafios com o desabastecimento na cadeia de suprimentos em um hospital público: relato de caso. *Revista de Gestão em Sistemas de Saúde*, v. 5, n. 1, p. 120–127, 2016. Doi: <https://doi.org/10.5585/rqss.v5i1.218>. Acesso em: 17 maio 2024.

LUO, Yuyu; LI, Guoliang; FAN, Ju; CHAI, Chengliang; TANG, Nan. Natural Language to SQL: State of the Art and Open Problems. *Proceedings of the VLDB Endowment*, v. 18, n. 12, p. 5466–5471, 2025. Doi: <https://dl.acm.org/doi/10.14778/3750601.3750696>. Acesso em: 17 maio 2024.

LUTKEVICH, Ben; COURTEMANCHE, Meredith. What is GitHub? *In: TECHTARGET [site]*, 9 jul. 2024. Disponível em: <https://www.techtarget.com/searchitoperations/definition/GitHub>. Acesso em: 7 jan. 2026.

MARIN, Nelly *et al.* (org.). *Assistência farmacêutica para gerentes municipais*. Rio de Janeiro: OPAS/MS, 2003.

MELO, Elaine Lopes de; OLIVEIRA, Luana de Souza. Farmácia hospitalar e o papel do farmacêutico no âmbito da assistência farmacêutica. *Revista JRG de Estudos Acadêmicos*, v. 4, n. 8, p. 287–299, 2021. Doi: <https://doi.org/10.5281/zenodo.4641016>. Acesso em: 17 maio 2024.

MONTAGNA, Sara *et al.* Data Decentralisation of LLM-Based Chatbot Systems in Chronic Disease Self-Management. *In: ACM CONFERENCE ON INFORMATION TECHNOLOGY FOR SOCIAL GOOD*, 2023. Lisboa. *Proceedings* [...]. Lisboa: ACM SIGCAS, p. 205–212.

NAIR, Uma S *et al.* Development of a Conversational Agent for Individuals Ambivalent About Quitting Smoking: Protocol for a Proof-of-Concept Study. *JMIR Research Protocols*, v. 12, n. 1, e44041, 2023. Doi: <https://doi.org/10.2196/44041>. Acesso em: 17 maio 2024.

NASCIMENTO NETO, Conrado Dias do *et al.* Inteligência artificial e novas tecnologias em saúde: desafios e perspectivas. *Brazilian Journal of Development*, v. 6, n. 2, p. 9431–9445, 2020. Doi: <https://doi.org/10.34117/bjdv6n2-306>. Acesso em: 31 maio 2024.

NIKOLIC, Irina A.; MAIKISCH, Harald. Public-Private Partnerships and Collaboration in the Health Sector: An Overview with Case Studies from Recent European Experience. Washington, DC: World Bank, 2006. (HNP Discussion Paper Series).

NUNES, Alexandre; MATOS, Andreia. Análise do desempenho dos hospitais em parceria público privada no Serviço Nacional de Saúde português. *Jornal Brasileiro de Economia da Saúde*, v. 9, n. 3, p. 304–309, 2017. Doi: <https://doi.org/10.21115/JBES.v9.n3.p304-9>. Acesso em: 14 mar. 2026.

PACKEISER, Priscila Becker; RESTA, Darielli Gindri. Farmacoeconomia: uma ferramenta para a gestão dos gastos com medicamentos em hospitais públicos. *Infarma – Ciências Farmacêuticas*, v. 26, n. 4, 2014. Doi: <https://doi.org/10.14450/2318-9312.v26.e4.a2014.pp215-223>. Acesso em: 31 maio 2024.

PEDROSO, Breno Carvalho; PEREIRA, Marluce Rofrigues; PEREIRA, Denilson Alves. Performance Evaluation of LLMs in the Text-to-SQL Task in Portuguese. *In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO (SBSI)*, 21., 2025, Recife, PE. *Anais* [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. Doi: 10.5753/sbsi.2025.246471. Disponível em: <https://sol.sbc.org.br/index.php/sbsi/article/view/34342>. Acesso em: 11 jan. 2026.

QWEN TEAM. *Qwen3 technical report*. *In: QWEN Team* [site], 15 maio 2025. Disponível em: <https://github.com/QwenLM/Qwen3>. Acesso em: 7 jan. 2026.

RAIMUNDO, Eliane Amâncio; DIAS, Cleidson Nogueira; GUERRA, Mariana. Logística de medicamentos e materiais em um hospital público do Distrito Federal. *Revista de Administração Hospitalar e Inovação em Saúde*, v. 12, n. 2, p. 61–69, 2015. Doi: <https://doi.org/10.21450/rahis.v12i2.2384>. Acesso em: 31 maio 2024.

RAZA, Muhammad Ahmer *et al.* Artificial Intelligence (AI) in Pharmacy: An Overview of Innovations. *Innovations in Pharmacy*, v. 13, n. 2, 2022. Doi: <https://doi.org/10.24926/iip.v13i2.4839>. Acesso em: 17 maio 2024.

- REIS, Amanda. *A inteligência artificial no setor de saúde: conceitos e aplicações*. Texto para discussão n. 99. São Paulo: Instituto de Estudos de Saúde Suplementar, 2023. Disponível em: <https://www.iess.org.br/sites/default/files/2023-12/TD%20-%2099.pdf>. Acesso em: 31 maio 2024.
- ROB, Peter; CORONEL, Carlos. *Sistemas de banco de dados: projeto, implementação e gerenciamento*. São Paulo: Pearson, 2010.
- ONG, Guoguang *et al.* Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*, v. 6, n. 3, p. 291–301, 2020. Doi: <https://doi.org/10.1016/j.eng.2019.08.015>. Acesso em: 17 maio 2024.
- SCHUHMACHER, Alexander *et al.* Big Techs and Startups in Pharmaceutical R&D. *Drug Discovery Today*, v. 26, n. 10, p. 2226–2231, 2021. Doi: <https://doi.org/10.1016/j.drudis.2021.04.028>. Acesso em: 17 maio 2024.
- SHEVAT, Amir. *Designing bots: creating conversational experiences*. Sebastopol: O'Reilly Media, 2017.
- SILVA, Letícia O.; SILVA, Paulo H. C.; SILVA, Fabrício A. Leis de escala para Text-to-SQL: um estudo sobre a relação entre tamanho e desempenho de modelos de linguagem. In: SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS (SBBBD), 40., 2025, Fortaleza, CE. *Anais [...]*. Fortaleza: SBC, 2025, p. 140–153. Doi: <https://doi.org/10.5753/sbbd.2025.247042>. Acesso em: 10 jan. 2026.
- SILVA, Luiza Pinheiro Alves da. *Orçamento temático de acesso a medicamentos 2019*. Brasília: INESC, 2020.
- SINGER, Philipp *et al.* H2O-Danube-1.8B Technical Report. *ArXiv*, 2024. Disponível em: <https://arxiv.org/abs/2401.16818>. Acesso em: 16 jan. 2026.
- SOARES, Ronan Pardo. QueryAugment-RAG: aprimorando a recuperação e geração de informações em modelos de linguagem de grande escala. *Revista Contemporânea*, v. 4, n. 6, p. 1–15, 2024. Doi: 10.56083/RCV4N6-156. Acesso em: 10 jan. 2026.
- SUMATHI, S.; ESAKKIRAJAN, S. *Fundamentals of Relational Database Management Systems*. New York: Springer, 2007.
- TACCA, Adriano; ROCHA, Leonel Severo. Inteligência artificial: reflexos no sistema do Direito. *Revista do Programa de Pós-Graduação em Direito da UFC*, v. 38, n. 2, p. 53–68, jul./dez. 2018.
- TEIXEIRA, Carlos Augusto Leal; ANDRADE, Leonardo Guimarães. O uso de indicadores de qualidade e desempenho para evitar custos e desperdícios de medicamentos na farmácia hospitalar. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, v. 8, n. 3, p. 1558–1566, 2022. Doi: <https://doi.org/10.51891/rease.v8i3.4740>. Acesso em: 17 maio 2024.
- TORCHIA, Mariateresa; CALABRÒ, Andrea; MORNER, Michèle. Public-Private Partnerships in the Health Care Sector: A Systematic Review of the Literature. *Public*

Management Review, v. 17, n. 2, p. 236–261, 2015. Doi:

<https://doi.org/10.1080/14719037.2013.792380>. Acesso em: 14 mar. 2026.

TUSTUMI, Francisco; ANDREOLLO, Nelson Adami; AGUILAR-NASCIMENTO, José Eduardo. Future of the Language Models in Healthcare: The Role of CHATGPT.

ABCD: Arquivos Brasileiros de Cirurgia Digestiva, v. 36, e1727, 2023. Doi:

<https://doi.org/10.1590/0102-672020230002e1727>. Acesso em: 17 maio 2024.

WILSON, Lee; MARASOIU, Mariana. The Development and Use of Chatbots in

Public Health: Scoping Review. *JMIR Human Factors*, v. 9, n. 4, e35882, 2022. Doi:

<https://doi.org/10.2196/35882>. Acesso em: 14 mar. 2026.

YANG, Rui *et al.* Retrieval-Augmented Generation for Generative Artificial Intelligence in Health Care. *npj Health Systems*, v. 2, n. 2, 2025. Doi:

10.1038/s44401-024-00004-1. Disponível em:

<https://www.nature.com/articles/s44401-024-00004-1.pdf>. Acesso em: 18 jan. 2026.

ZHOU, Yongchao *et al.* Large Language Models are Human-Level Prompt

Engineers. Published as a conference paper at ICLR 2023. *arXiv*, 2023. Disponível

em: <https://arxiv.org/pdf/2211.01910>. Acesso em: 17 jan. 2026.

APÊNDICE A – Estrutura do Formulário (Google Forms – Fase Exploratória – Avaliação das Demandas)

Prezado(a) participante,

Este formulário faz parte de um estudo cujo objetivo é aplicar o uso de Processamento de Linguagem Natural (PLN) para otimização da gestão farmacêutica em um hospital de grande porte. Para isto, pretende-se desenvolver um *chatbot*.

Chatbots são programas de computador projetados para imitar a conversa humana por meio de interações de texto ou voz. Eles podem variar desde programas rudimentares com respostas pré-programadas até sistemas sofisticados que aproveitam a inteligência artificial (IA) para Processamento de Linguagem Natural, ou seja, eles entendem o que você diz e aprendem com cada conversa, ficando cada vez melhores!

A sua participação é fundamental para identificar as demandas dos gestores e aprimorar o processo de gestão de insumos farmacêuticos. As informações fornecidas serão confidenciais e utilizadas exclusivamente para fins de pesquisa.

Obrigado pela sua colaboração.

Atenciosamente,
Equipe de Pesquisa

Seção 1: Dados de Caracterização

1.1. Nome Completo: _____

1.2. Idade: _____

1.3 Sexo:

- Masculino
- Feminino
- Prefiro não responder

1.4 Cargo Atual: _____

1.5 Tempo de Experiência na Área Farmacêutica ou relacionadas: _____

Seção 2: Conhecimento Autorreferido sobre Assistência Farmacêutica

2.1 Como você avalia seu conhecimento sobre assistência farmacêutica no âmbito hospitalar?

- Muito bom
- Bom
- Regular
- Ruim
- Muito ruim

2.2 Você já participou de treinamentos ou capacitações relacionadas à gestão de estoques na farmácia hospitalar?

- Sim
- Não

Seção 3: Informações Consultadas Rotineiramente

3.1 Quais são as principais informações consultadas rotineiramente para amparar o processo de gestão de estoques? (Você pode selecionar mais de uma opção)

- Quantidade de medicamentos em estoque
- Data de validade dos medicamentos
- Histórico de consumo dos medicamentos
- Previsão de demanda de medicamentos
- Fornecedores e prazos de entrega
- Custos
- Outros (especificar): _____

3.2 Qual a frequência com que você consulta essas informações?

- Diariamente
- Semanalmente
- Mensalmente
- Raramente
- Nunca

3.3 Qual é a principal dificuldade enfrentada para consulta dessas informações?

3.4 Quais ferramentas ou sistemas você utiliza atualmente para acessar essas informações?

3.5 Em sua opinião, como um *chatbot* poderia facilitar o acesso a essas informações?

Seção 4: Avaliação Geral

4.1 Você estaria disposto a participar de uma fase de testes do *chatbot* desenvolvido no projeto?

- Sim
- Não

4.2 Deixe aqui seus comentários ou sugestões adicionais:

Agradecemos imensamente sua participação e colaboração nesta pesquisa. Sua contribuição é fundamental para o sucesso do nosso projeto.

A Equipe de Pesquisa

APÊNDICE B – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO (TCLE)

Título do Estudo: **Uso de Processamento de Linguagem Natural para otimização da gestão farmacêutica em um hospital de grande porte público**

Pesquisador Responsável: Raphael Augusto Teixeira de Aguiar

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Prezado(a) Senhor(a),

Gostaríamos de convidá-lo(a) a participar da pesquisa intitulada “**Uso de Processamento de Linguagem Natural para otimização da gestão farmacêutica em um hospital de grande porte público**”. Esta pesquisa está sendo conduzida por Fabíola de Oliveira Simões, discente do Curso de Mestrado Profissional em Gestão de Serviços de Saúde, sob orientação do Prof. Dr. Raphael Augusto Teixeira de Aguiar.

Objetivo da Pesquisa

O objetivo principal desta pesquisa é desenvolver um protótipo de *chatbot* para a gestão farmacêutica hospitalar, visando facilitar o acesso a informações relacionadas à gestão de estoque pelos farmacêuticos. Especificamente, pretendemos identificar as demandas dos gestores, elaborar contextos de conversação para o *chatbot* e avaliar a interação do sistema com os usuários.

Metodologia da Pesquisa

A pesquisa será realizada no Hospital Metropolitano Doutor Célio de Castro e os participantes serão envolvidos em duas fases principais:

a) Fase de avaliação das demandas:

Nesta fase, você será abordado(a) para preencher um questionário online via Google Forms. Essa atividade tem como objetivo identificar as demandas e necessidades específicas que o *chatbot* deve atender na gestão farmacêutica hospitalar. Durante esta fase, você será solicitado(a) a descrever suas tarefas diárias e desafios enfrentados na gestão farmacêutica.

b) Fase de avaliação do protótipo:

Após o desenvolvimento inicial do protótipo do *chatbot*, você será convidado(a) a participar da fase de avaliação do protótipo. Nessa etapa, será solicitado que interaja com a ferramenta, realizando questionamentos que fazem parte da rotina na gestão da farmácia hospitalar. Esta fase incluirá sessões práticas no hospital, onde você utilizará o *chatbot* em situações simuladas de gestão farmacêutica. Cada sessão terá a duração aproximada de trinta minutos. Seu *feedback* será coletado através de novo questionário ajudando a ajustar e melhorar o *chatbot*.

Os pesquisadores estarão disponíveis para fornecer acompanhamento e assistência durante todas as atividades.

Consulta de Dados

Será solicitada sua autorização para consultar dados específicos relacionados ao uso do *chatbot*, tais como *logs* de conversa e *feedback* fornecido.

Riscos e Desconfortos

Os riscos associados à privacidade dos dados são mitigados por meio de estratégias rigorosas de anonimização e segurança dos dados. Como plano de contingência, se problemas forem identificados durante a avaliação inicial ou as fases de teste, ajustes serão feitos tanto no treinamento dos modelos quanto nas práticas de manuseio de dados.

Benefícios

A participação nesta pesquisa pode trazer benefícios como a melhoria dos serviços farmacêuticos no hospital, além de contribuir para avanços na área de gestão farmacêutica hospitalar.

Voluntariedade

A sua participação nesta pesquisa é totalmente voluntária. Você tem o direito de recusar ou desistir da participação a qualquer momento, sem qualquer prejuízo ou penalidade.

Sigilo e Confidencialidade

Os dados coletados serão mantidos em sigilo e confidencialidade. Os resultados da pesquisa serão utilizados em trabalhos científicos e apresentações sem revelar sua identidade.

Ressarcimento e Indenização

Você não terá qualquer despesa para participar desta pesquisa e não receberá remuneração por sua participação. Caso ocorra algum dano relacionado à pesquisa, você terá direito à indenização.

Armazenamento e Responsabilidade pelos Dados

Este Termo de Consentimento Livre e Esclarecido será assinado em duas vias antes do início do estudo. Uma via será entregue a você, enquanto a outra será armazenada em formato físico, em local seguro e com acesso restrito aos colaboradores do setor de Farmácia, nas dependências do Hospital Metropolitano Doutor Célio de Castro.

Os formulários online serão preenchidos via Google Forms, e as respostas serão armazenadas em um ambiente protegido e acesso restrito. Os dados estarão seguros em servidores do Google, que cumprem padrões internacionais de segurança da informação, além de atenderem à Lei Geral de Proteção de Dados (LGPD).

A responsabilidade pela proteção e confidencialidade dos documentos físicos e dos dados armazenados digitalmente será do pesquisador principal Raphael Augusto Teixeira de Aguiar e da discente do Curso de Mestrado Profissional em Gestão de Serviços de Saúde, Fabíola de Oliveira Simões. Os dados permanecerão armazenados por um período de 5 (cinco) anos após a conclusão da pesquisa. Ao término desse período, todos os documentos e informações serão destruídos e descartados de maneira segura, garantindo o sigilo e a privacidade dos participantes.

Contato para Esclarecimentos

Caso você tenha qualquer dúvida geral sobre o estudo ou sobre a sua participação, o pesquisador responsável está à disposição para esclarecimentos. Por favor, entre em contato.

Pesquisador Principal:

Nome: Raphael Augusto Teixeira de Aguiar

Telefone: (31) XXXXX-XXXX⁴

Email: raphael.aguiar@gmail.com

Mestranda:

Nome: Fabíola de Oliveira Simões

⁴ No documento original, constavam números de telefone da equipe, os quais foram omitidos nesta versão por se tratar de publicação pública.

Telefone: (31) X XXXX-XXXX

Email: fabiola.simoese@hotmail.com

Se você tiver questões relacionadas às questões éticas do estudo, ou se sentir que seus direitos como participante estão sendo violados, você pode entrar em contato com o Comitê de Ética em Pesquisa.

Comissão de Ética em Pesquisa (CEP-UFMG):

Telefone: 3409-4592

Email: coep@prpq.ufmg.br

Endereço:

Av. Presidente Antônio Carlos, 6627, Pampulha – Belo Horizonte – MG –

CEP 31270-901 – Unidade Administrativa II – 2º Andar – Sala: 2005

Telefone: (031) 3409-4592 – E-mail: coep@prpq.ufmg.br

Horário de atendimento: 09:00 às 11:00 / 14:00 às 16:00

Declaração de Consentimento

Eu, _____, após ter lido e compreendido as informações acima, concordo em participar voluntariamente desta pesquisa. Recebi uma via deste termo para meus registros.

Participante:

Assinatura: _____

Data: ____/____/____

Rubrica: _____

Pesquisador Principal:

Assinatura: _____

Data: ____/____/____

Rubrica: _____

APÊNDICE C – TERMO DE COMPROMISSO DE UTILIZAÇÃO DE DADOS (TCUD)

1. Identificação dos membros do grupo de pesquisa

Nome completo (sem abreviação)	RG	Assinatura
Raphael Augusto Teixeira de Aguiar	MG-X.XXX.XXX ⁵	
Fabíola de Oliveira Simões	MG-XX.XXX.XXX	

2. Identificação da pesquisa

- a) Título do Projeto: **Uso de Processamento de Linguagem Natural para otimização da gestão Farmacêutica em um hospital de grande porte**
- b) Departamento/Faculdade/Curso: **Departamento de Medicina Preventiva e Social (DMPS) / Faculdade de Medicina / Mestrado Profissional em Gestão de Serviços de Saúde (MPGSS)**
- c) Pesquisador Responsável: **Raphael Augusto Teixeira de Aguiar**

3. Descrição dos Dados

São dados a serem coletados somente após aprovação do projeto de pesquisa pelo Comitê de Ética da Universidade Federal de Minas Gerais (CEP-UFMG):

Os dados serão coletados para treinamento de um modelo que, por meio de Processamento de Linguagem Natural, apoiará os farmacêuticos na gestão de estoque e acompanhamento de indicadores da qualidade das atividades desenvolvidas pelo serviço de farmácia hospitalar. A ferramenta auxiliará na análise de todas as movimentações de estoque dos insumos sob responsabilidade da farmácia (materiais médicos, medicamentos, saneantes, OPME – Órteses, Próteses e Materiais Especiais), além de informações relacionadas aos custos dispensados à estes. Para tal, será necessário acesso à base de dados (ORACLE) do sistema de gestão hospitalar Tasy (somente outbound/saída) por meio de internet ou a disponibilização de um arquivo .csv com dados extraídos da base.

⁵ No documento original, constavam dados reais da equipe, os quais foram omitidos nesta versão por se tratar de publicação pública.

Nome completo (sem abreviação)	Assinatura
Raphael Augusto Teixeira de Aguiar	
Fabíola de Oliveira Simões	

5. Autorização da Instituição

Declaramos, para os devidos fins, que cederemos aos pesquisadores apresentados neste termo o acesso aos dados solicitados para serem utilizados nesta pesquisa.

Esta autorização está condicionada ao cumprimento do(a) pesquisador(a) aos requisitos da Resolução nº 466/12 e suas complementares, comprometendo-se o(a) mesmo(a) a utilizar os dados dos participantes da pesquisa, exclusivamente para os fins científicos, mantendo o sigilo e garantindo a não utilização das informações em prejuízo das pessoas e/ou das comunidades.

Antes de iniciar a coleta de dados o(a) pesquisador(a) deverá apresentar o projeto completo e o Parecer Consubstanciado devidamente aprovado, emitido por Comitê de Ética em Pesquisa Envolvendo Seres Humanos, credenciado ao Sistema CEP/CONEP.

Belo Horizonte, ____ de _____ de ____

Nome legível/assinatura e carimbo do responsável pela anuência da Instituição

APÊNDICE D – CARACTERIZAÇÃO DA AMOSTRA (FASE DE PESQUISA E PLANEJAMENTO)

CARACTERIZAÇÃO DA AMOSTRA

A idade dos respondentes variou entre 25 e 44 anos, apresentando mediana de 35 anos e média de 33,7 anos. Em relação ao sexo, 77,8% dos participantes identificaram-se como feminino e 22,2% como masculino.

Em relação à ocupação, como esperado, observou-se predominância de funções farmacêuticas, sendo que 6 dos 9 participantes (66,7%) se enquadraram na categoria Farmacêutico Hospitalar e Clínico. As demais funções foram: Coordenadora de Assistência Farmacêutica (11,1%), Enfermeira da Qualidade de Materiais (11,1%), Diretora de Recursos Terapêuticos, Diagnóstico e Segurança Assistencial (11,1%) e Assessora da diretoria (11,1%).

O tempo de experiência na área apresentou amplitude de 0,5 a 15 anos (considerando “Inferior a 6 meses” como 0,5 ano), com mediana de 4 anos e média de 6,28 anos.

CONHECIMENTO AUTORREFERIDO SOBRE ASSISTÊNCIA FARMACÊUTICA

No que se refere à autopercepção do conhecimento em assistência farmacêutica hospitalar, a maior parte dos participantes classificou seu nível de conhecimento como “bom” (66,7%), enquanto 22,2% o avaliaram como “muito bom” e 11,1% como “regular”.

Em relação à participação prévia em treinamentos ou capacitações voltadas à gestão de estoques, 77,8% dos respondentes já haviam participado de atividades formativas nesta área, ao passo que 22,2% relataram não ter tido essa experiência.

ANEXO – PRODUTO TÉCNICO

**UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENFERMAGEM
PROGRAMA DE PÓS-GRADUAÇÃO EM
GESTÃO DE SERVIÇOS DE SAÚDE
MESTRADO PROFISSIONAL EM GESTÃO DE SERVIÇOS DE SAÚDE**

Fabíola de Oliveira Simões

**PRODUTO TÉCNICO/TECNOLÓGICO RESULTANTE DO TRABALHO “Aplicação
de Processamento de Linguagem Natural como ferramenta para suporte à Gestão
Farmacêutica em ambiente hospitalar”**

Belo Horizonte
2026

Fabíola de Oliveira Simões

PRODUTO TÉCNICO/TECNOLÓGICO RESULTANTE DO TRABALHO “Aplicação de Processamento de Linguagem Natural como ferramenta para suporte à Gestão Farmacêutica em ambiente hospitalar”

Produto técnico/tecnológico resultante do Trabalho de Conclusão de Mestrado apresentado ao curso de Mestrado Profissional do Programa de Pós-Graduação em Gestão de Serviços de Saúde, da Escola de Enfermagem da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Mestre em Gestão de Serviços de Saúde.

Área de concentração: Gestão de Serviços de Saúde.

Linha de pesquisa: Tecnologias Gerenciais em Saúde

Orientador: Prof. Dr. Raphael Augusto Teixeira de Aguiar

Coorientador: Prof. Dr. Leonardo Chaves Dutra da Rocha

RESUMO

A gestão de farmácias hospitalares envolve volumes significativos de dados e requer acesso rápido a informações confiáveis para apoiar os processos de tomada de decisão. Este estudo metodológico teve como objetivo explorar a aplicação do Processamento de Linguagem Natural (PLN) na gestão de farmácias hospitalares por meio do desenvolvimento de um protótipo de *chatbot* baseado em um Modelo Amplo de Linguagem (LLM) da família Qwen. A pesquisa foi conduzida em duas fases complementares: a primeira fase de caráter exploratório, dedicada à identificação das principais solicitações de informação relacionadas à gestão de estoques e uma segunda fase de desenvolvimento e avaliação funcional do protótipo, implementado em Python. A avaliação da ferramenta focou em um conjunto estruturado de perguntas representativas das necessidades comuns dos profissionais, resultantes da análise das respostas dos participantes durante a primeira fase. Os resultados mostraram uma taxa de sucesso geral de execução de 85%, indicando que o modelo foi capaz de gerar consultas SQL válidas na maioria dos casos. O desempenho mostrou-se satisfatório para perguntas simples de baixa a moderada complexidade, que produziram respostas consistentes e acionáveis. No entanto, uma proporção significativa de consultas mais complexas não levou a respostas semanticamente corretas, apesar da execução técnica bem-sucedida, evidenciando principalmente limitações na interpretação de regras de negócio implícitas. De modo geral, o protótipo demonstrou viabilidade técnica e potencial para aplicação prática, ao mesmo tempo que ressaltou a necessidade de ajustes metodológicos e refinamentos do modelo para melhorar a precisão das respostas em cenários de tomada de decisão mais complexos.

Palavras-chave: Processamento de Linguagem Natural; Sistemas Inteligentes; Serviço de Farmácia Hospitalar; Inteligência Artificial.

ABSTRACT

Hospital pharmacy management involves significant volumes of data and requires rapid access to reliable information to support decision-making processes. This methodological study aimed to explore the application of Natural Language Processing (NLP) in hospital pharmacy management through the development of a chatbot prototype based on a Large Language Model (LLM) from the Qwen family. The research was conducted in two complementary phases: the first, exploratory phase, dedicated to identifying the main information requests related to inventory management, and a second phase of development and functional evaluation of the prototype, implemented in Python. The evaluation of the tool focused on a structured set of questions representative of the common needs of professionals, resulting from the analysis of participants' responses during the first phase. The results showed an overall execution success rate of 85%, indicating that the model was able to generate valid SQL queries in most cases. Performance proved satisfactory for simple questions of low to moderate complexity, which produced consistent and actionable responses. However, a significant proportion of more complex queries did not lead to semantically correct answers, despite successful technical execution, mainly highlighting limitations in the interpretation of implicit business rules. Overall, the prototype demonstrated technical feasibility and potential for practical application, while also emphasizing the need for methodological adjustments and model refinements to improve the accuracy of responses in more complex decision-making scenarios.

Keywords: Natural Language Processing; Intelligent Systems; Pharmacy Service, Hospital; Artificial Intelligence.

LISTA DE ILUSTRAÇÕES

Figura 1 – Fluxo simplificado.....	100
Figura 2 – Esquema de funcionamento do sistema RAG.....	101
Figura 3 – Como o sistema RAG funciona	102
Quadro 1 – Síntese das métricas de avaliação	106

LISTA DE ABREVIATURAS E SIGLAS

LLM	Large Language Models (Modelos Amplos de Linguagem)
RAG	Recuperação Aumentada por Geração
SQL	Structured Query Language (Linguagem de Consulta Estruturada)
TI	Tecnologia da Informação

SUMÁRIO

1 INTRODUÇÃO	96
2 OBJETIVO DO PRODUTO TÉCNICO	97
3 CARACTERIZAÇÃO DO PROTÓTIPO DESENVOLVIDO	98
3.1 Arquitetura e funcionamento do sistema	98
3.2 Conversão de linguagem natural em SQL.....	99
3.3 Sistema de Recuperação Aumentada por Geração (RAG)	100
3.3.1 Modelo de inteligência artificial utilizado.....	102
3.3.2 Parâmetros configuráveis do sistema.....	103
3.3.3 Estrutura do banco de dados	103
3.3.4 Vantagens da abordagem proposta	104
4 MÉTRICAS DE AVALIAÇÃO DO MODELO.....	106
5 DISPONIBILIZAÇÃO E REPRODUTIBILIDADE DO PROTÓTIPO	109
6 CONSIDERAÇÕES FINAIS	110
REFERÊNCIAS.....	111

RELATÓRIO TÉCNICO

PROTÓTIPO DE CONVERSÃO DE LINGUAGEM NATURAL EM SQL COM USO DE RECUPERAÇÃO AUMENTADA POR GERAÇÃO (RAG)

1 INTRODUÇÃO

A crescente digitalização dos serviços de saúde tem ampliado de forma significativa o volume de dados armazenados em sistemas de informação hospitalares (Yang *et al.*, 2025). Embora esses dados sejam fundamentais para o apoio à gestão, à tomada de decisão e ao planejamento assistencial, seu acesso permanece, em grande parte, restrito a profissionais com conhecimento técnico específico em linguagens de consulta a bancos de dados, como o SQL (Linguagem de Consulta Estruturada). Esse cenário impõe barreiras à utilização plena das informações disponíveis, especialmente por gestores e profissionais da área da saúde que não possuem formação em tecnologia da informação.

No contexto da gestão farmacêutica hospitalar, essa limitação torna-se ainda mais sensível, uma vez que decisões relacionadas a estoque, consumo, cobertura de medicamentos e materiais impactam diretamente a segurança do paciente, a eficiência operacional e os custos institucionais (Raimundo; Dias; Guerra, 2015; Lima; Câmara, 2016; Bruns; Luiza; Oliveira, 2014). Diante desse desafio, tecnologias baseadas em Processamento de Linguagem Natural e modelos de linguagem de grande porte surgem como alternativas para aproximar usuários não técnicos dos sistemas de dados estruturados.

Este relatório técnico apresenta o protótipo Text2SQL desenvolvido como produto técnico associado a dissertação de mestrado profissional intitulada “Aplicação de Processamento de Linguagem Natural como ferramenta para suporte à Gestão Farmacêutica em ambiente hospitalar”, alinhado às diretrizes de produção técnica aplicadas a cursos *stricto sensu* de natureza profissional (CAPES, 2019). O protótipo tem como finalidade permitir que perguntas formuladas em linguagem natural sejam automaticamente convertidas em consultas SQL executáveis, com apoio de uma arquitetura baseada em Recuperação Aumentada por Geração (Retrieval-Augmented Generation – RAG).

2 OBJETIVO DO PRODUTO TÉCNICO

O objetivo geral deste produto técnico é apresentar e documentar o desenvolvimento de um protótipo de sistema capaz de converter perguntas em linguagem natural para consultas SQL, utilizando Modelo Amplo de Linguagem (LLM) apoiado por um mecanismo de recuperação de contexto estruturado.

Como objetivos específicos, destacam-se:

- Descrever a arquitetura técnica do protótipo Text2SQL;
- Explicar o funcionamento do sistema de Recuperação Aumentada por Geração aplicado ao contexto de bancos de dados hospitalares;
- Apresentar as principais funcionalidades implementadas no protótipo;
- Documentar as métricas utilizadas para avaliação do desempenho do modelo, em nível conceitual;
- Disponibilizar informações que favoreçam a reprodutibilidade e a evolução futura da solução proposta.

3 CARACTERIZAÇÃO DO PROTÓTIPO DESENVOLVIDO

O protótipo Text2SQL consiste em uma aplicação computacional desenvolvida em linguagem Python, projetada para permitir a interação entre usuários e bases de dados relacionais por meio de perguntas formuladas em linguagem natural (Silva; Silva; Silva, 2025). O sistema foi concebido para operar sobre bancos de dados PostgreSQL e utilizou o LLM Qwen3-32B-AWQ, por sua natureza *Open Source*, capacidade de compreensão semântica em língua portuguesa e desempenho satisfatório em tarefas de geração de texto estruturado.

O público-alvo do protótipo compreende gestores e profissionais da área da saúde que necessitam acessar informações armazenadas em bancos de dados hospitalares, mas que não possuem domínio técnico da linguagem SQL. O protótipo foi desenvolvido com foco no domínio da gestão farmacêutica hospitalar, considerando estruturas de dados relacionadas a estoque, cadastro de materiais e medicamentos, movimentações e indicadores de cobertura.

Do ponto de vista funcional, o sistema permite a geração automática de consultas SQL, a execução dessas consultas no banco de dados, a recuperação dos resultados e a avaliação da qualidade das respostas geradas, além de disponibilizar uma interface web para visualização e análise dos resultados.

3.1 Arquitetura e funcionamento do sistema

O protótipo Text2SQL foi desenvolvido a partir de uma organização que distribui suas funcionalidades em partes bem definidas, cada uma responsável por uma etapa do processo de transformação das perguntas em informações úteis para o usuário. Essa forma de organização permite que o sistema funcione de maneira integrada e facilita sua manutenção e adaptação a diferentes realidades institucionais.

De modo geral, o sistema recebe perguntas formuladas em linguagem natural (texto em português), identifica quais informações do banco de dados são necessárias para respondê-las e realiza automaticamente a consulta correspondente. Todo esse processo ocorre de forma transparente para o usuário, que não precisa conhecer a estrutura do banco de dados nem dominar a linguagem SQL para obter os resultados desejados.

O acesso ao protótipo pode ocorrer tanto por meio de uma interface de linha de comando quanto por uma interface web (Shiny). A primeira é voltada principalmente

ao uso técnico e à execução de testes, enquanto a interface web foi pensada para facilitar a interação de usuários não especialistas. Independentemente da forma de acesso, o funcionamento interno do sistema permanece o mesmo, assegurando que as respostas geradas sejam consistentes.

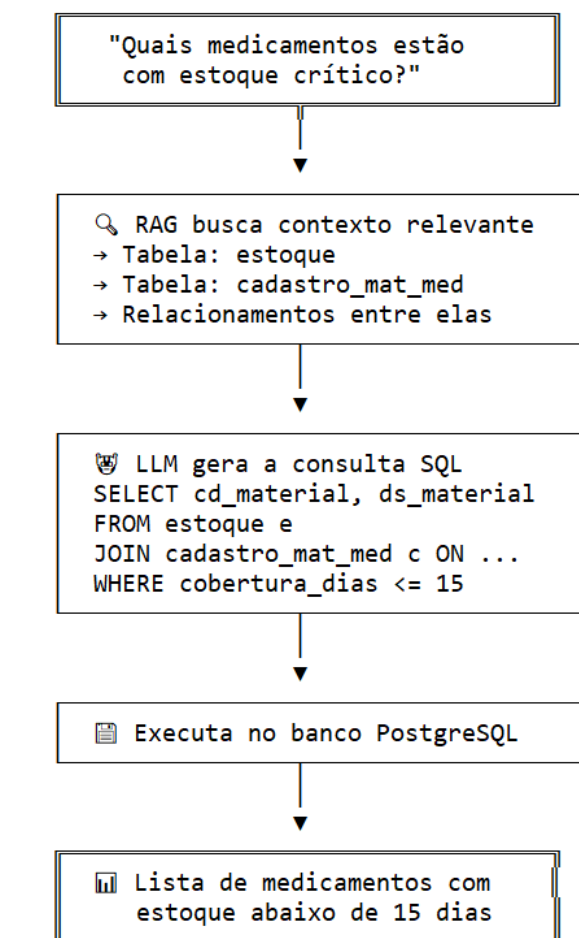
Após o envio da pergunta, o sistema realiza o processamento necessário, executa a consulta no banco de dados e apresenta os resultados de forma organizada. Dessa maneira, o protótipo atua como um intermediário entre o usuário e o banco de dados, simplificando o acesso à informação e ampliando o uso dos dados disponíveis para apoio à gestão.

3.2 Conversão de linguagem natural em SQL

O núcleo do protótipo é o mecanismo de conversão de linguagem natural em SQL, realizado por meio de um LLM. Esse modelo recebe como entrada a pergunta formulada pelo usuário em português e é instruído a produzir uma consulta SQL correspondente, respeitando a estrutura do banco de dados e regras de negócio previamente definidas.

Entretanto, a utilização direta de modelos de linguagem para geração de SQL apresenta limitações importantes, especialmente quando o modelo não dispõe de informações suficientes sobre o esquema do banco de dados, como nomes de tabelas, colunas e relacionamentos. Para mitigar essas limitações, o protótipo incorpora um mecanismo de Recuperação Aumentada por Geração (RAG), abordagem que tem se mostrado eficaz para aumentar a confiabilidade e a rastreabilidade das respostas geradas por modelos de linguagem (Soares, 2024; Yang *et al.*, 2025). A Figura 1 esquematiza o processo de construção da ferramenta:

Figura 1 – Fluxo simplificado

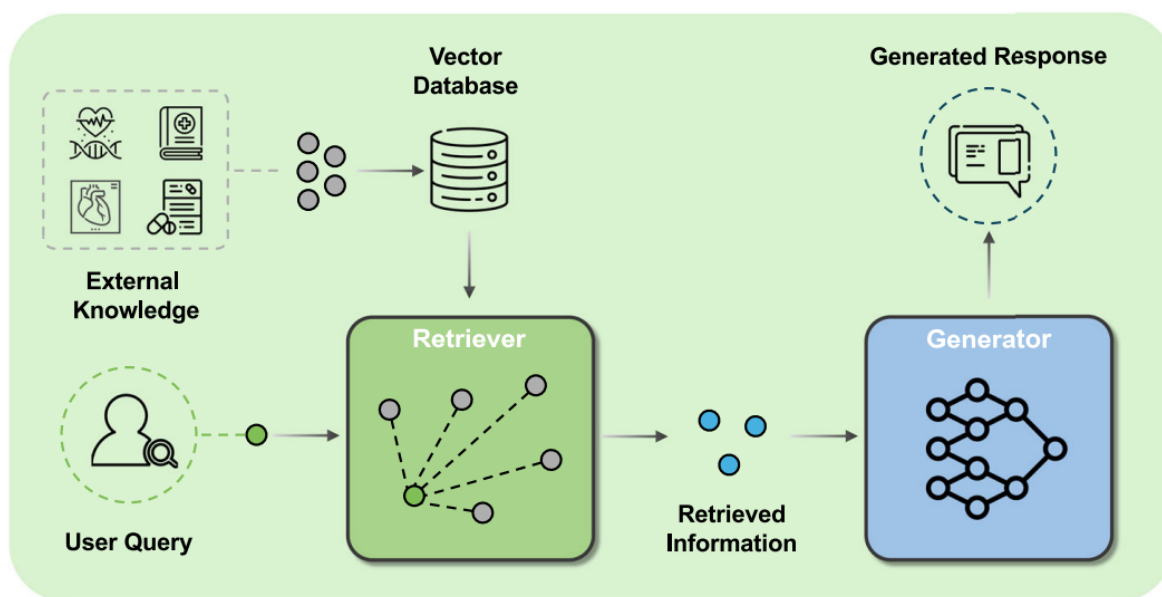


Fonte: Elaborado pela equipe de TI responsável pelo desenvolvimento do protótipo.

3.3 Sistema de Recuperação Aumentada por Geração (RAG)

A técnica de Recuperação Aumentada por Geração é responsável por fornecer ao modelo de linguagem informações contextuais relevantes sobre o banco de dados antes da geração da consulta SQL, seguindo a arquitetura clássica de indexação, recuperação e geração descrita na literatura (Yang *et al.*, 2025). A Figura 2 ilustra o funcionamento geral dessa abordagem. No modelo apresentado, a pergunta do usuário é inicialmente utilizada para recuperar informações relevantes a partir de uma base de conhecimento externa. Em seguida, esse conjunto de informações é incorporado ao processo de geração da resposta pelo LLM.

Figura 2 – Esquema de funcionamento do sistema RAG



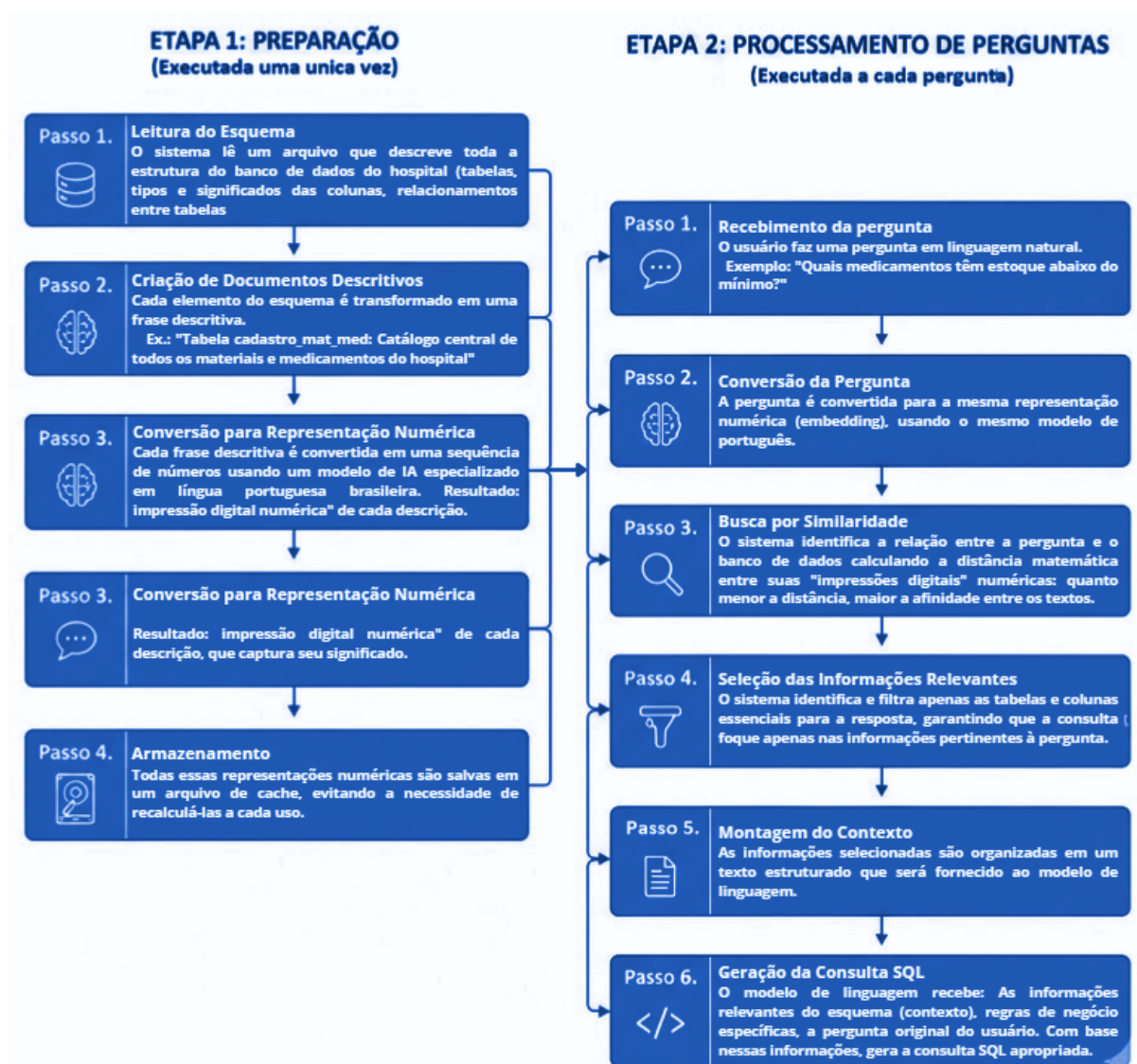
Fonte: Yang *et al.* (2025).

Na fase de indexação, essas informações são transformadas em representações vetoriais por meio de modelos de *embeddings* treinados para o idioma português. Os vetores gerados são armazenados em um índice vetorial, permitindo a recuperação eficiente de informações com base em similaridade semântica. *Embeddings* são representações numéricas de textos. Imagine que cada palavra ou frase é convertida em uma “coordenada” em um espaço matemático multidimensional. Palavras com significados semelhantes ficam “próximas” nesse espaço. Por exemplo: as palavras “medicamento”, “remédio” e “fármaco” teriam coordenadas muito próximas entre si, enquanto a palavra “cadeira” ficaria distante. Isso permite que o computador entenda que textos são semanticamente similares, mesmo usando palavras diferentes. Quando o usuário submete uma pergunta em linguagem natural, o sistema gera um vetor correspondente à pergunta e realiza uma busca por similaridade no índice vetorial.

Como resultado, são recuperados apenas os elementos do esquema do banco de dados mais relevantes para aquela pergunta específica. Esse conjunto de informações é então incorporado ao *prompt* fornecido ao modelo de linguagem, aumentando a probabilidade de geração de uma consulta SQL correta e coerente. A Figura 3 esquematiza este fluxo no protótipo desenvolvido.

Essa abordagem permite reduzir ambiguidades, aumentar a confiabilidade das consultas geradas e melhorar a transparência do processo de geração, uma vez que o contexto utilizado pelo modelo pode ser explicitamente identificado.

Figura 3 – Como o sistema RAG funciona



Fonte: Elaborado pela autora.

3.3.1 Modelo de inteligência artificial utilizado

Embora o modelo de linguagem responsável pela geração das consultas SQL seja o Qwen3-32B-AWQ, o sistema utiliza um segundo modelo, o BERTimbau (neuralmind/bert-large-portuguese-cased),⁶ para a etapa de recuperação de contexto. Esse modelo é empregado para converter textos em representações numéricas,

⁶ <https://huggingface.co/neuralmind/bert-large-portuguese-cased>

permitindo comparar semanticamente a pergunta do usuário com descrições do esquema do banco de dados. Dessa forma, o BERTimbau atua exclusivamente na seleção das informações relevantes, enquanto o Qwen é responsável pela geração final da consulta SQL.

O modelo BERTimbau foi escolhido por:

- Especialização em português: Foi treinado especificamente com textos em português brasileiro, compreendendo melhor expressões, gírias e estruturas gramaticais do nosso idioma.
- Tamanho e capacidade: Possui 335 milhões de parâmetros ajustáveis, o que lhe confere alta capacidade de compreensão textual.
- Sensibilidade a maiúsculas: Diferencia “Medicamento” de “medicamento”, preservando informações importantes presentes em nomes técnicos.
- Treinamento abrangente: Foi treinado em um grande volume de textos brasileiros, incluindo artigos, notícias e documentos técnicos.

3.3.2 Parâmetros configuráveis do sistema

O sistema permite ajustar diversos parâmetros para otimizar seu funcionamento:

- Limiar de similaridade (`similarity_threshold`): Define quão “parecido” um resultado precisa ser para ser considerado relevante. Valor padrão: 0.3 (em uma escala de 0 a 1);
- Número máximo de tabelas (`max_tables`): Quantas tabelas podem ser incluídas no contexto. Valor padrão: 5 tabelas;
- Colunas por tabela (`max_columns_per_table`): Quantas colunas de cada tabela são incluídas. Valor padrão: 10 colunas;
- Tamanho máximo do contexto (`max_context_length`): Limite de caracteres do contexto gerado. Valor padrão: 32.768 caracteres.

3.3.3 Estrutura do banco de dados

O esquema do banco de dados foi documentado em um formato estruturado (YAML) contendo, para cada tabela:

- Descrição da tabela: Explicação do propósito e conteúdo da tabela
Exemplo: “Catálogo central de todos os materiais/medicamentos do hospital”
- Lista de colunas: Cada coluna com seu tipo de dado e descrição
Exemplo: “`cd_material` (número inteiro) – Código único do item”

- Chave primária: Identificador único de cada registro na tabela
- Chaves estrangeiras: Relacionamentos com outras tabelas

Exemplo: “A tabela de estoque se relaciona com o cadastro de materiais através do código do material”

- Índices: Colunas otimizadas para buscas rápidas

Esta documentação detalhada permite que o sistema compreenda não apenas a estrutura técnica, mas também o significado de cada elemento no contexto hospitalar.

3.3.4 Vantagens da abordagem proposta

A adoção de uma arquitetura baseada em RAG apresenta um conjunto de vantagens relevantes no contexto de sistemas que integram modelos de linguagem e bases de dados estruturadas. Essas vantagens dizem respeito tanto ao desempenho computacional quanto à qualidade das respostas geradas e à possibilidade de aplicação do sistema em diferentes contextos institucionais.

a) Economia de recursos computacionais

A abordagem proposta evita o envio integral da documentação do banco de dados ao modelo de linguagem a cada solicitação. Em vez disso, apenas as informações consideradas relevantes para a pergunta formulada pelo usuário são selecionadas e fornecidas como contexto. Essa estratégia contribui para a redução do volume de dados processados, resultando em menor tempo de resposta e em uso mais eficiente dos recursos computacionais, aspecto especialmente importante em ambientes que demandam escalabilidade e controle de custos.

b) Aumento da precisão das respostas

Ao receber um conjunto de informações direcionado e alinhado à pergunta realizada, o modelo de linguagem opera com menor risco de ambiguidades. A restrição do contexto às tabelas, colunas e relacionamentos pertinentes reduz a probabilidade de geração de consultas incorretas ou inconsistentes, favorecendo respostas mais precisas e compatíveis com a estrutura real do banco de dados.

c) Escalabilidade do sistema

A arquitetura adotada permite que o sistema mantenha desempenho satisfatório mesmo em bancos de dados de grande porte, com elevado número de

tabelas e atributos. A busca por similaridade semântica, utilizada para a recuperação do contexto relevante, apresenta custo computacional reduzido quando comparada à análise exaustiva de todo o esquema do banco, tornando a solução adequada para cenários complexos e em expansão.

d) Compreensão semântica das perguntas

Diferentemente de abordagens baseadas exclusivamente em correspondência literal de termos, o sistema considera o significado das palavras utilizadas na pergunta. Dessa forma, expressões semanticamente relacionadas são reconhecidas como equivalentes, mesmo quando não há coincidência exata de vocabulário. Isso permite, por exemplo, que perguntas sobre “remédios vencidos” sejam associadas a informações relacionadas a medicamentos e prazos de validade, ampliando a capacidade de interpretação do sistema.

e) Adaptabilidade a diferentes contextos

A solução proposta apresenta elevada flexibilidade, uma vez que sua adaptação a novos bancos de dados depende, essencialmente, da atualização do arquivo de esquema que descreve a estrutura dos dados. Essa característica dispensa alterações no código-fonte principal do sistema, facilitando sua reutilização em diferentes instituições e contextos de aplicação, o que é particularmente relevante em ambientes organizacionais heterogêneos.

4 MÉTRICAS DE AVALIAÇÃO DO MODELO

Como etapa fundamental do desenvolvimento do protótipo, foi construída uma base de referência denominada *ground truth*, composta por consultas SQL elaboradas e validadas pela equipe de Tecnologia da Informação (TI). O *ground truth* corresponde a um conjunto de consultas consideradas corretas e representativas das perguntas de interesse, servindo como padrão para comparação e avaliação das consultas geradas automaticamente pelo modelo.

A avaliação de modelos em Machine Learning é um aspecto crítico que não pode ser negligenciado. Para garantir que esses modelos sejam capazes de explicar e prever fenômenos com precisão, é muito importante utilizar as métricas corretas. O desempenho de cada pergunta do conjunto de avaliação foi analisado por meio de métricas quantitativas para avaliação de sistemas de recuperação de informação e geração automática de consultas, tais como **acurácia**, **precisão**, **recall** e **F1-score** (Quadro 1). Essas métricas foram calculadas a partir da comparação entre as consultas SQL geradas automaticamente pelo modelo e a base de referência (*ground truth*) previamente validada.

Quadro 1 – Síntese das métricas de avaliação

Métrica	Descrição	Fórmula
Acurácia	Indica uma performance geral do modelo. Dentre todas as classificações, quantas o modelo classificou corretamente.	$\frac{VP + VN}{VP + FV + FP + FN}$
Precisão	Dentre todas as classificações da classe positiva que o modelo fez, quantas estão corretas.	$\frac{VP}{VP + FP}$
<i>Recall</i> <i>/ Revocação / Sensibilidade:</i>	Dentre todas as classificações da classe positiva como valor esperado, quantas estão corretas.	$\frac{VP}{VP + FN}$
<i>F1-Score</i>	Média harmônica entre precisão e <i>Recall</i> .	$\frac{2 * Precisão * Recall}{Precisão + Recall}$

Fonte: Pinho; Gaspar; Sassi (2024).

a) Acurácia

A acurácia representa a proporção geral de acertos do sistema, considerando tanto os itens corretamente retornados quanto os itens incorretamente incluídos, devendo ser interpretada com cautela em tarefas de recuperação de informação. A fórmula da acurácia é a proporção de previsões corretas em relação ao total de

previsões realizadas. Para entender os acertos, utiliza-se a matriz de confusão,⁷ que separa Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). Embora seja amplamente utilizada, sua interpretação deve ser realizada com cautela em problemas de recuperação de informação, sobretudo quando há desbalanceamento entre respostas corretas e incorretas (ex: 99% spam, 1% não-spam), o modelo pode ter alta acurácia (99%) sem realmente aprender a identificar a classe minoritária) (Pinho; Gaspar; Sassi, 2024; Ferri; Hernández-Orallo; Modroi, 2009).

b) Precisão

A precisão é uma métrica que indica a proporção de itens retornados pelo sistema que estão corretos em relação ao total de itens apresentados como resposta, sendo amplamente utilizada na avaliação de sistemas de recuperação de informação e aprendizado de máquina. A fórmula da precisão é o número de verdadeiros positivos dividido pela soma dos verdadeiros positivos e falsos positivos. É particularmente importante quando uma previsão positiva seria dispendiosa se se revelasse errada (Pinho; Gaspar; Sassi, 2024; Ferri; Hernández-Orallo; Modroi, 2009).

No contexto de perguntas que resultam em listas de registros, a precisão permite avaliar o grau de exatidão das respostas fornecidas pelo modelo.

c) Recall

O *recall* mede a proporção de itens corretos que foram efetivamente recuperados pelo sistema em relação ao total de itens esperados, refletindo a capacidade do modelo em não omitir respostas relevantes. Essa métrica é particularmente relevante em contextos nos quais a omissão de informações pode comprometer a tomada de decisão (Pinho; Gaspar; Sassi, 2024; Ferri; Hernández-Orallo; Modroi, 2009).

⁷ **Verdadeiro Positivo** corresponde às situações em que o valor esperado é positivo e o modelo identifica corretamente esse caso, retornando um resultado também positivo, o que caracteriza um acerto da predição realizada.

Verdadeiro Negativo refere-se aos casos em que o resultado esperado é negativo e o modelo corretamente retorna um valor negativo, caracterizando uma predição correta.

Falso Positivo ocorre quando o modelo indica um resultado positivo para um caso cujo valor real é negativo, caracterizando uma classificação incorreta.

Falso Negativo ocorre quando o modelo classifica um caso como negativo, apesar de o valor real ser positivo, caracterizando uma predição incorreta (Pinho; Gaspar; Sassi, 2024).

A fórmula para *recall* é o número de verdadeiros positivos dividido pela soma de verdadeiros positivos e falsos negativos. Uma alta pontuação de *recall* mostra que o modelo está capturando a maioria dos casos positivos reais nos dados.

d) *F1-score*

O *F1-score* corresponde à média harmônica entre precisão e *recall*, sendo obtido pelo dobro do produto dessas métricas dividido pela soma de ambas. Essa medida é particularmente relevante em contextos nos quais tanto os falsos positivos quanto os falsos negativos possuem impacto significativo, oferecendo uma avaliação mais equilibrada do desempenho do modelo, especialmente em conjuntos de dados desbalanceados (Pinho; Gaspar; Sassi, 2024; Ferri; Hernández-Orallo; Modroi, 2009).

5 DISPONIBILIZAÇÃO E REPRODUTIBILIDADE DO PROTÓTIPO

Com o objetivo de promover a transparência, a reprodutibilidade científica e a possibilidade de evolução contínua da solução proposta, o código-fonte do protótipo Text2SQL foi disponibilizado em um repositório GitHub, prática recomendada para produtos técnicos e pesquisas aplicadas em ciência de dados e saúde digital. O repositório contém a estrutura completa do projeto, incluindo os módulos de geração de consultas, o sistema de recuperação de contexto, os scripts de avaliação e a interface web.

O acesso ao repositório permite que outros pesquisadores e profissionais da área da saúde compreendam a implementação do protótipo, adaptem-no a diferentes contextos institucionais e contribuam para seu aprimoramento futuro. Adicionalmente, eventuais melhorias de desempenho ou ajustes no protótipo serão registradas no GitHub, permitindo o acompanhamento da evolução do projeto. É importante destacar que os resultados apresentados refletem o estágio atual do desenvolvimento, podendo não ser definitivos, já que fatores como o tempo disponível para execução e a conclusão para apresentação à banca final podem influenciar o desempenho observado.

O Link para acesso ao repositório GitHub: <https://github.com/fabiolaoliveira-2026/Projeto-Prot-tipo-Gest-o-Farmac-utica.git>

6 CONSIDERAÇÕES FINAIS

Este relatório técnico apresentou o protótipo Text2SQL desenvolvido como produto técnico de um trabalho de conclusão de mestrado, com foco na conversão de linguagem natural em consultas SQL apoiada por um mecanismo de Recuperação Aumentada por Geração. O documento descreveu a arquitetura do sistema, seu funcionamento, os principais componentes técnicos e as métricas utilizadas para avaliação do desempenho do modelo.

O protótipo proposto demonstra o potencial de aplicação de LLMs e técnicas de RAG no apoio à gestão farmacêutica hospitalar, contribuindo para a ampliação do acesso a informações estratégicas por usuários não técnicos. Como produto técnico, a solução apresentada constitui uma base para futuras evoluções, adaptações e estudos aplicados no contexto dos sistemas de informação em saúde.

REFERÊNCIAS

- BRUNS, Suelma de Fátima; LUIZA, Vera Lucia; OLIVEIRA, Egléubia Andrade de. Gestão da assistência farmacêutica em municípios do estado da Paraíba (PB): olhando a aplicação de recursos públicos. *Revista de Administração Pública*, v. 48, n. 3, p. 745–765, 2014. Doi: <https://doi.org/10.1590/0034-76121502>. Acesso em: 3 jan. 2026.
- COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR (CAPES). *Produção técnica*: Grupo de Trabalho. Brasília: CAPES, 2019. Disponível em: <https://www.gov.br/capes/pt-br/centrais-de-conteudo/10062019-producao-tecnica-pdf/view>. Acesso em: 18 jan. 2026.
- FERRI, C.; HEERNÁNDEZ-ORALLO, J.; MODROIU, R. An Experimental Comparison of Performance Measures for Classification. *Pattern Recognition Letters*, v. 30, n. 1, p. 27–38, 2009. Doi: <https://doi.org/10.1016/j.patrec.2008.08.010>. Acesso em: 10 jan. 2026.
- LIMA, Rodrigo Fonseca; CÂMARA, Tássio Fontes Moreira. Desafios com o desabastecimento na cadeia de suprimentos em um hospital público: relato de caso. *Revista de Gestão em Sistemas de Saúde*, v. 5, n. 1, p. 120–127, 2016. Doi: <https://doi.org/10.5585/rgss.v5i1.218>. Acesso em: 10 jan. 2026.
- PINHO, Cintia Maria de Araújo; GASPARG, Marcos Antonio; SASSI, Renato José. Aplicação de técnicas de inteligência artificial para classificação de fuga ao tema em redações. *Educação em Revista*, Belo Horizonte, v. 40, e39773, 2024. Doi: <http://dx.doi.org/10.1590/0102-469839773>. Disponível em: <https://www.scielo.br/j/edur/a/YPPVRRKGdLdFbXTQXMmZgCc/?format=html&lang=pt>. Acesso em: 18 jan. 2026.
- RAIMUNDO, Eliane Amâncio; DIAS, Cleidson Nogueira; GUERRA, Mariana. Logística de medicamentos e materiais em um hospital público do Distrito Federal. *Revista de Administração Hospitalar e Inovação em Saúde*, v. 12, n. 2, p. 61–69, 2015. Doi: <https://doi.org/10.21450/rahis.v12i2.2384>. Acesso em: 3 jan. 2026.
- SOARES, Ronan Pardo. QueryAugment-RAG: aprimorando a recuperação e geração de informações em modelos de linguagem de grande escala. *Revista Contemporânea*, v. 4, n. 6, p. 1–15, 2024. Doi: 10.56083/RCV4N6-156. Acesso em: 3 jan. 2026.
- YANG, Rui et al. Retrieval-Augmented Generation for Generative Artificial Intelligence in Health Care. *npj Health Systems*, v. 2, n. 2, 2025. Doi: 10.1038/s44401-024-00004-1. Disponível em: <https://www.nature.com/articles/s44401-024-00004-1.pdf>. Acesso em: 18 jan. 2026.