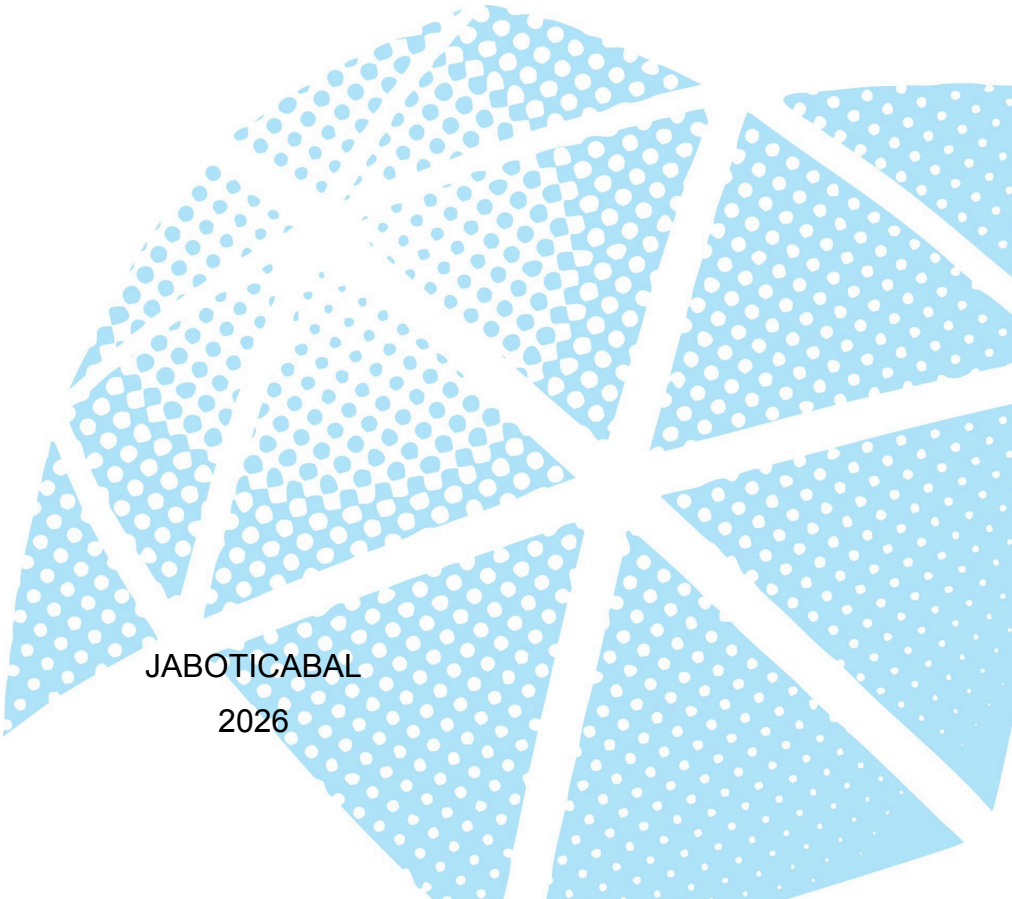
UNIVERSIDADE ESTADUAL PAULISTA - UNESP

Faculdade de Ciências Agrárias e Veterinárias - Câmpus de Jaboticabal

INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA ANÁLISE DO CONSUMO DE COMBUSTÍVEL DE CAMINHÕES NO TRANSPORTE DE CANA-DE-AÇÚCAR

Lucas Eduardo Zonfrilli



JABOTICABAL

2026

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
Faculdade de Ciências Agrárias e Veterinárias - Câmpus de Jaboticabal

INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA ANÁLISE DO CONSUMO DE COMBUSTÍVEL DE CAMINHÕES NO TRANSPORTE DE CANA-DE-AÇÚCAR

Discente: Lucas Eduardo Zonfrilli

Orientadora: Profa. Dra. Tatiana Fernanda Canata

Coorientador: Prof. Dr. Almir Rogério Camolesi

Dissertação apresentada à Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal, para obtenção do título de Mestre em Agronomia (Produção Vegetal).
Área de Concentração: Produção Vegetal

JABOTICABAL
2026

Z87i

Zonfrilli, Lucas Eduardo

Inteligência artificial generativa para análise do consumo de combustível de caminhões no transporte de cana-de-açúcar / Lucas Eduardo Zonfrilli. -- Jaboticabal, 2026

61 p. : il., tabs.

Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientadora: Tatiana Fernanda Canata

Coorientador: Almir Rogério Camolesi

1. Cana-de-açúcar. 2. Inteligência artificial. 3. SQL (Linguagem de programação de computador). 4. Transporte rodoviário. 5. Caminhões. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

IMPACTO POTENCIAL DESTA PESQUISA

Sistema de inteligência artificial que responde, por mensagem no celular, a perguntas de supervisores de frota sobre o consumo de diesel de caminhões canavieiros, com acurácia de pelo menos 99% e custo de R$ 0,07 por consulta, apoiando decisões que reduzem custos e emissões de CO_2 no setor sucroenergético.

POTENTIAL IMPACT OF THIS RESEARCH

An artificial intelligence system that answers, via mobile messaging, fleet supervisors' questions about the diesel consumption of sugarcane trucks, with at least 99% accuracy at R$ 0.07 per query, supporting decisions that reduce costs and CO_2 emissions in the sugar-energy sector.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: INTELIGÊNCIA ARTIFICIAL GENERATIVA PARA ANÁLISE DO CONSUMO DE COMBUSTÍVEL DE CAMINHÕES NO TRANSPORTE DE CANA-DE-AÇÚCAR

AUTOR: LUCAS EDUARDO ZONFRILLI
ORIENTADORA: TATIANA FERNANDA CANATA
COORIENTADOR: ALMIR ROGERIO CAMOLESI

Aprovado como parte das exigências para obtenção do Título de Mestre em Agronomia (Produção Vegetal), pela Comissão Examinadora:

Prof. Dr. ALMIR ROGERIO CAMOLESI (Participaçao Presencial)
Departamento Ciência da Computação / Fundação Educacional do Município de Assis (FEMA) - Assis/SP

gov.br Documento assinado digitalmente
FABIO EDER CARDOSO
Data: 24/07/2026 16:16:45-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. FÁBIO EDER CARDOSO (Participaçao Virtual)
Departamento: CERPII - Centro Referencial de Propriedade Intelectural e Inovação em Ciência da Informação / Fundação Educacional do Município de Assis (FEMA) - Assis/SP

Prof. Dr. CARLOS EDUARDO ANGELI FURLANI (Participaçao Presencial)
Departamento de Engenharia / UNESP / Câmpus de Jaboticabal - FCAV

Jaboticabal, 24 de julho de 2026.

Dedico este trabalho à memória do meu pai Carlos, um homem de fé, agricultor e muito trabalhador, que me incentivou a buscar o conhecimento.

AGRADECIMENTOS

À Consciência Amorosa Universal, que nos guia e toca nossos corações em busca daquilo em que verdadeiramente acreditamos, nos ajudando a encontrar nosso lugar na roda, neste mundo cheio de possibilidades, tendo em vista que este é o despertar mais genuíno do ser humano.

À minha família, que foi base, apoio e motivação para que tudo isto se tornasse realidade: à minha mãe Lucimeire, que é uma excelente professora, a quem admiro e tenho como exemplo e referência, sendo meu incentivo e inspiração para buscar o conhecimento como fonte de progresso; ao meu irmão Carlos, com quem construí uma parceria ímpar de apoio, mansidão e direcionamento; ele torna isso possível, pois sei que sempre terei com quem contar; bem como ao meu sobrinho, tão amado e querido.

À minha companheira de vida, Gabriele, pelas horas dedicadas a me escutar sobre as descobertas durante o processo, pela resiliência diante da distância e pelos cuidados comigo e pela revisão do português deste trabalho. Amo todos vocês.

À Profa. Dra. Tatiana Fernanda Canata, que sempre esteve atenta aos detalhes e prazos, me ajudando a construir esta jornada. Agradeço pela oportunidade de fazer parte do Grupo de Geoprocessamento e Agricultura de Precisão da Unesp de Jaboticabal (GEOAP) e pela orientação, que foi de grande valia.

Ao coorientador Prof. Dr. Almir Rogério Camolesi, pela troca de ideias e pelo incentivo constante para que eu pudesse evoluir e entregar o melhor trabalho possível.

Aos membros da banca examinadora, Prof. Dr. Fábio Éder Cardoso e Prof. Dr. Carlos Eduardo Angeli Furlani, pela leitura atenta e pelas contribuições a este trabalho.

Agradeço também ao meu conselheiro nesta caminhada, Darío Palmieri, por todas as conversas francas e pelo direcionamento no momento certo.

De modo especial, agradeço a todos os colegas de trabalho do Laboratório de Máquinas e Mecanização Agrícola (LAMMA): todos os momentos compartilhados e as experiências vividas ajudaram-me a lapidar conhecimentos, expandir fronteiras e a evoluir como ser humano.

Agradeço ao Prof. Dr. Rouverson Pereira da Silva, por oportunizar essa vivência com um time tão forte, e também por todas as trocas durante esta caminhada.

Aos amigos da República Independência, que me receberam e me proporcionaram viver um lado mais família da UNESP, assim como à Dona Cláudia, por todo o carinho e cuidado.

Agradeço à CAPES pela concessão da bolsa de estudos que viabilizou a realização desta pesquisa. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

RESUMO

O transporte rodoviário de cana-de-açúcar concentra uma parcela expressiva dos custos e das emissões do setor sucroenergético, e a telemetria embarcada gera elevado volume de dados cuja conversão em informação útil ao supervisor de frota ainda constitui um gargalo. Este trabalho teve por objetivo desenvolver e avaliar um sistema multiagente de inteligência artificial generativa, estruturado conforme o ciclo Planejar, Fazer, Checar e Agir (PDCA) e disponibilizado via mensageria, para converter perguntas em linguagem natural em linguagem de consulta estruturada (SQL) a partir de dados de telemetria do transporte canavieiro. A arquitetura sem servidor (*serverless*), executada integralmente em nuvem, associou quatro agentes especializados às fases Planejar, Fazer, Checar e Agir, com Geração Aumentada por Recuperação (RAG) e modelo de linguagem *Gemini 2.5 Flash*. Sob delineamento fatorial 2×2 (geração sem exemplos, *zero-shot*, e geração com exemplos recuperados, *few-shot*, cruzada com conjuntos de 900 e de 732 perguntas), foram conduzidas 3.264 execuções a partir de dados reais de 30.791 viagens. A acurácia de execução variou de 99,00% a 99,59%, sem ganho estatisticamente significativo da RAG pelo teste de McNemar, o que caracterizou o efeito de teto (*ceiling effect*), embora a RAG tenha alterado a distribuição dos erros entre os níveis de complexidade. O custo variável foi de R$ 0,0651 por pergunta, e o estimador de custo em tempo real apresentou desvio de apenas −2,5% frente à fatura real da nuvem, com ponto de paridade entre custo fixo e variável em aproximadamente 507 consultas por dia. A pegada de carbono por consulta, no cenário ajustado por energia livre de carbono (*carbon-free energy*), situou-se entre 0,77 mg e 0,85 mg de CO_2 por consulta, e não se registraram erros de vinculação ao esquema nem bloqueios de conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD) durante todas as execuções. Os resultados indicam potencial de economia e eficiência de custo, além de viabilidade para a adoção do sistema como ferramenta auditável de apoio à decisão no transporte de cana-de-açúcar.

Palavras-chave: cana-de-açúcar; PDCA; RAG; sistema multiagente; telemetria agrícola; *text-to-SQL*.

ABSTRACT

Road transport of sugarcane accounts for a significant share of the costs and emissions of the sugar-energy sector, and onboard telemetry generates a data volume whose conversion into useful information for the fleet supervisor remains a bottleneck. This work aimed to develop and evaluate a multi-agent generative artificial intelligence system, based on the Plan, Do, Check and Act (PDCA) cycle and delivered through messaging, to convert natural-language questions into Structured Query Language (SQL) queries over sugarcane transport telemetry data. The serverless architecture, running entirely in the cloud, associated four specialized agents with the Plan, Do, Check and Act phases, with Retrieval-Augmented Generation (RAG) and the Gemini 2.5 Flash language model. Under a 2×2 factorial design (zero-shot vs. few-shot generation, crossed with sets of 900 and 732 questions), 3,264 executions were carried out over 30,791 real trips. Execution accuracy ranged from 99.00% to 99.59%, with no statistically significant gain from RAG by the McNemar test, characterizing a ceiling effect, although RAG changed the error distribution across complexity levels. The variable cost was R$ 0.0651 per question, and the real-time cost estimator showed a deviation of only −2.5% against the actual cloud bill, with an operational break-even point at approximately 507 queries per day. The carbon footprint adjusted for carbon-free energy ranged from 0.77 to 0.85 mg CO_2 per query, and zero schema-linking errors and zero blocks by the Brazilian General Data Protection Law (LGPD) compliance layer were recorded across all executions. The results indicate potential for cost savings and cost efficiency, as well as strong feasibility for adopting the system as an auditable decision-support tool in sugarcane transport.

Keywords: sugarcane; PDCA; RAG; multi-agent system; agricultural telemetry; text-to-SQL.

LISTA DE ILUSTRAÇÕES

Figura 1 – Fluxo operacional do transporte de cana-de-açúcar, da área agrícola à agroindústria, e interface conversacional do sistema multiagente via mensageria *Telegram* 22

Figura 2 – Comparação conceitual entre a arquitetura de agente único e a arquitetura multiagente, evidenciando a decomposição da tarefa complexa em agentes especializados 23

Figura 3 – Mapeamento dos quatro agentes especializados às fases do ciclo PDCA (Planejar, Fazer, Checar e Agir), com os respectivos papéis e o alinhamento à norma ISO/IEC 42001 24

Figura 4 – Arquitetura *serverless* do sistema multiagente na plataforma de nuvem *Google Cloud Platform* 25

Figura 5 – Exemplos de resposta do agente PDCA na interface do *Telegram*, com previsão de eficiência por modelo de caminhão e o mecanismo de validação humana (certo ou errado) 28

Figura 6 – Interface de controle do agente *quest-runner* no *Telegram*, demonstrando a identificação da bateria de avaliação, o conjunto de perguntas em execução, a estimativa de duração e os comandos de gestão do experimento 33

Figura 7 – Síntese comparativa das quatro condições experimentais: (a) acurácia de execução; (b) custo variável por pergunta; (c) custo variável total; (d) duração total do experimento; (e) média de *tokens* por consulta; (f) erros por taxonomia 38

Figura 8 – Estrutura econômica do sistema: (a) custo por pergunta estimado pelo sistema versus fatura real, com composição de custos fixos e variáveis; (b) ponto de paridade entre custo fixo e variável; (c) cascata de custo do modelo de linguagem por componente de cobrança; (d) cascata de volume de *tokens* 40

Figura 9 – Validação do estimador de custo contra a fatura real, por serviço de nuvem, com a separação das parcelas variáveis (dependentes do volume de consultas) e fixas 42

Figura 10 – Projeção do custo anual da safra (204 dias operacionais) por frequência de consulta, demonstrando a composição entre custos fixos e variáveis em três cenários de carga 43

Figura 11 – Latência do ciclo PDCA: (a) tempo médio de processamento por agente nas quatro condições experimentais; (b) latência mediana por nível de complexidade da consulta (simples, detalhada e complexa) 45

Figura 12 – Desempenho do componente de Geração Aumentada por 46

Recuperação: (a) custo da RAG frente aos demais componentes; (b) impacto na acurácia e cobertura por conjunto; (c) latência por agente, com a recuperação representando 0,7% do ciclo; (d) distribuição da similaridade semântica e limiares de recuperação

Figura 13 – Pegada de carbono da inferência nas quatro condições experimentais: (a) emissões de CO_2 totais acumuladas nos cenários bruto e ajustado por energia livre de carbono; (b) emissão média de CO_2 por consulta; (c) comparação entre as abordagens *zero-shot* e *few-shot* 47

Figura 1A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível simples 57

Figura 2A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível detalhado 58

Figura 3A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível complexo 58

LISTA DE TABELAS

Tabela 1 – Importância das variáveis preditoras do modelo de consumo de combustível (*AutoML* do *BigQuery*), obtida por inteligência artificial explicável (*Explainable AI*), expressa em valor de atribuição absoluto e em importância relativa (proporção da atribuição total) 27

Tabela 2 – Acurácia de execução por condição experimental, com intervalos de confiança de 95% pelo método de Wilson 37

Tabela 3 – Distribuição de erros por tipo, segundo a taxonomia de Yu *et al.* (2018) e Lei *et al.* (2025), e por condição experimental 48

LISTA DE ABREVIATURAS E SIGLAS

CTTA	Corte, Transbordo, Transporte e Administração da cana-de-açúcar
FS	Geração com exemplos recuperados (*few-shot*)
GCP	Plataforma de nuvem *Google Cloud Platform*
IC	Intervalo de confiança
IP	Protocolo de Internet (*Internet Protocol*)
ISO/IEC 42001	Norma internacional de sistema de gestão de inteligência artificial
LGPD	Lei Geral de Proteção de Dados Pessoais
LLM	Modelo de linguagem de grande escala (*Large Language Model*)
MAE	Erro médio absoluto (*Mean Absolute Error*)
PDCA	Planejar, Fazer, Checar e Agir (*Plan, Do, Check, Act*)
PUE	Eficiência no uso de energia do centro de dados (*Power Usage Effectiveness*)
RAG	Geração Aumentada por Recuperação (*Retrieval-Augmented Generation*)
RMSE	Raiz do erro quadrático médio (*Root Mean Squared Error*)
SKU	Unidade de manutenção de estoque (*Stock Keeping Unit*)
SLM	Modelo de linguagem de pequeno porte (*Small Language Model*)
SQL	Linguagem de consulta estruturada (*Structured Query Language*)
ZS	Geração sem exemplos (*zero-shot*)

SUMÁRIO

1	**INTRODUÇÃO**	**14**
2	**REVISÃO DE LITERATURA**	**17**
2.1	AGRICULTURA DE PRECISÃO E TELEMETRIA NO TRANSPORTE CANAVIEIRO	17
2.2	MODELOS DE LINGUAGEM DE GRANDE ESCALA E CONVERSÃO DE LINGUAGEM NATURAL EM LINGUAGEM DE CONSULTA ESTRUTURADA (*TEXT-TO-SQL*)	17
2.3	GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG)	18
2.4	CUSTO COMPUTACIONAL, SUSTENTABILIDADE E CONFORMIDADE LGPD	19
2.5	CICLO PDCA E GOVERNANÇA DE INTELIGÊNCIA ARTIFICIAL (ISO/IEC 42001)	20
3	**MATERIAL E MÉTODOS**	**21**
3.1	FONTE E NATUREZA DOS DADOS	21
3.2	ARQUITETURA DO SISTEMA MULTIAGENTE	22
3.3	MODELO DE PREDIÇÃO DO CONSUMO DE COMBUSTÍVEL	25
3.4	DELINEAMENTO EXPERIMENTAL	28
3.4.1	Ambiente de execução automatizada e orquestração do experimento	31
3.4.2	Base de exemplos validados e recuperação vetorial	33
3.4.3	Estratificação por complexidade da consulta SQL	34
3.5	MÉTRICAS E ANÁLISE ESTATÍSTICA	35
4	**RESULTADOS E DISCUSSÃO**	**37**
4.1	ACURÁCIA DE EXECUÇÃO E IMPACTO DA RAG	37
4.2	CUSTO OPERACIONAL E VALIDAÇÃO CONTRA A FATURA REAL	39
4.3	LATÊNCIA DO CICLO PDCA	43
4.4	DESEMPENHO DA GERAÇÃO AUMENTADA POR RECUPERAÇÃO	45
4.5	PEGADA DE CARBONO DA INFERÊNCIA	47
4.6	CATEGORIZAÇÃO DE ERROS E CONFORMIDADE COM A LGPD	48
4.7	SÍNTESE, LIMITAÇÕES E DIREÇÕES FUTURAS	49
5	**CONCLUSÃO**	**51**
	REFERÊNCIAS	**52**
	APÊNDICE A – EXEMPLOS DE CONSULTAS POR NÍVEL DE COMPLEXIDADE	**57**
	DADOS CURRICULARES	**59**

1 INTRODUÇÃO

O Brasil responde por cerca de 39% da produção mundial de cana-de-açúcar (Food and Agriculture Organization of the United Nations, 2024) e moeu cerca de 621 milhões de toneladas de colmos na safra de 2024, com produtividade média de 77,78 t ha^{-1} (Pecege, 2026). Sustentar essa escala depende de uma operação logística intensa e contínua: o ciclo de Corte, Transbordo, Transporte e Administração da cana-de-açúcar (CTTA) movimenta a matéria-prima da área agrícola até a agroindústria por meio de comboios rodoviários de alta capacidade (rodotrens e treminhões), que percorrem distâncias variáveis entre os talhões e a unidade industrial, parte por estradas rurais, parte por rodovias. Somente o custo de diesel do transporte da cana colhida atinge R$ 4,10 por tonelada (Pecege, 2026), o que representa dispêndio setorial da ordem de R$ 2,5 bilhões e emissão estimada superior a 1,12 bilhão de quilogramas de CO_2 por safra, considerando o fator de 2,68 kg CO_2 L^{-1} (Intergovernmental Panel on Climate Change, 2006). É no transporte, portanto, que pequenas variações de eficiência se convertem em grandes diferenças de custo e de emissões: Bartholomeu, Péra e Caixeta-Filho (2016) estimaram que um ganho de 0,5 km L^{-1} na frota reduziria em cerca de 20% as emissões de gases de efeito estufa do transporte rodoviário de cargas.

A despeito do volume expressivo de dados de telemetria embarcada em máquinas e caminhões, a transformação desses dados em informação acessível e em tempo real permanece como desafio no setor sucroenergético (Tyrychtr; Ulman; Vostrovský, 2015; Wolfert *et al.*, 2017). A literatura identifica que as principais lacunas estão mais ligadas à cultura analítica e aos fluxos de processamento de dados do que à infraestrutura física disponível (Kamilaris; Kartakoullis; Prenafeta-Boldú, 2017).

Na prática de campo, o supervisor de frota convive com um paradoxo: a operação gera mais dados do que ele consegue consultar. Em uma safra típica de 204 dias, cerca de nove supervisores, distribuídos por turnos, gerenciam dezenas de milhares de viagens. A consulta manual a planilhas e painéis exige um tempo incompatível com a agilidade necessária à beira do talhão; o dado existe, mas não chega a quem opera no exato momento em que a decisão precisa ser tomada.

Esse gargalo deixou de ser tecnológico ou estrutural. A pesquisa TIC Domicílios 2025 indica que 81% dos domicílios rurais brasileiros possuem acesso à Internet. Além disso, 92% dos usuários de Internet enviaram mensagens instantâneas nos três meses anteriores à coleta, sendo o envio por *WhatsApp* ou *Telegram* reportado por 86% dos usuários de telefone celular. Entre os usuários de inteligência artificial generativa, 50% declararam empregá-la em atividades profissionais (Comitê Gestor da Internet no Brasil, 2026).

O problema passa, então, a ser o de transformar dados de máquinas em informação contextualizada e entregá-la diretamente ao tomador de decisão, no canal de comunicação que ele já utiliza. Sistemas multiagentes baseados em modelos de linguagem de grande escala (LLMs) emergem nesse contexto como uma abordagem capaz de decompor tarefas complexas em subtarefas especializadas (Guo *et al.*, 2024), com aplicações já demonstradas no setor de energia (Gamage *et al.*, 2024), educação (Yan *et al.*, 2024), manutenção industrial (Kiangala; Wang, 2024) e agricultura (Ray, 2024; Yang *et al.*, 2025).

O componente cognitivo central desta pesquisa é a conversão de perguntas em linguagem natural (*natural language*, NL) para linguagem de consulta estruturada (*text-to-SQL*), tarefa que permite ao LLM acessar os bancos de dados estruturados de telemetria.

Modelos recentes alcançam acurácia superior a 80% em *benchmarks* abertos (Gao, D. *et al.*, 2024; Li *et al.*, 2023), porém apresentam degradação em esquemas complexos e domínios especializados não cobertos no treinamento (Lei *et al.*, 2025; Yu *et al.*, 2018). A Geração Aumentada por Recuperação (RAG) tem se consolidado como estratégia para mitigar essas limitações ao injetar exemplos validados, recuperados por similaridade vetorial, no contexto do modelo (Gao, Y. *et al.*, 2024; Lewis *et al.*, 2020).

A operação de LLMs em escala, contudo, impõe restrições de custo que motivaram o conceito de custo por resposta correta (Chen; Zaharia; Zou, 2023). A esse condicionante econômico somam-se exigências de segurança, de conformidade com princípios de governança de dados e de atendimento à Lei Geral de Proteção de Dados Pessoais (LGPD) (Brasil, 2018), requisitos incontornáveis para sistemas que processam dados operacionais e pessoais de operadores.

Complementarmente, o ciclo PDCA (Planejar, Fazer, Checar e Agir) fornece arcabouço de melhoria contínua com eficácia documentada para indicadores

agrícolas (Silva *et al.*, 2008; Voltarelli *et al.*, 2018) e alinha-se às diretrizes da norma ISO/IEC 42001 para sistemas de gestão auditáveis em inteligência artificial (Thiers; Harned, 2024).

Diante desse cenário, a pergunta de pesquisa que orienta este trabalho é: em que medida um sistema multiagente de inteligência artificial generativa, orquestrado segundo o ciclo PDCA, é capaz de converter perguntas em linguagem natural em consultas SQL corretas sobre dados de telemetria do transporte canavieiro, com baixo custo e em conformidade com a LGPD, e qual o ganho marginal da Geração Aumentada por Recuperação neste contexto? A hipótese central é a de que a RAG, ao recuperar exemplos validados, eleva significativamente a acurácia da execução em relação à geração de respostas sem utilizar exemplos.

Para responder a essa questão, o objetivo geral foi desenvolver e avaliar um sistema multiagente de inteligência artificial generativa capaz de responder, por mensageria instantânea, a perguntas do cotidiano de supervisores de frota sobre dados de telemetria do transporte de cana-de-açúcar, considerando acurácia das respostas, consumo de *tokens*, latência e estabilidade em execução prolongada.

Como objetivos específicos, definiram-se: (i) projetar e implementar a arquitetura *serverless* multiagente em nuvem, mapeando os quatro agentes às fases Planejar, Fazer, Checar e Agir; (ii) quantificar a acurácia de execução e o efeito marginal da RAG sob delineamento fatorial 2×2, cruzando a estratégia de engenharia de *prompt*, em dois níveis (geração sem exemplos, *zero-shot*, e geração com exemplos recuperados, *few-shot*), com a composição do conjunto de perguntas; (iii) validar um estimador de custo de execução em tempo real contra faturamento efetivo do provedor de nuvem e determinar o volume de consultas necessário para diluir o custo fixo do sistema; e (iv) avaliar a pegada de carbono da inferência e a conformidade da arquitetura proposta com a LGPD.

2 REVISÃO DE LITERATURA

2.1 AGRICULTURA DE PRECISÃO E TELEMETRIA NO TRANSPORTE CANAVIEIRO

A agricultura de precisão consolidou-se como um paradigma de gestão geolocalizada, fundamentado na coleta intensiva de dados, bem como na sua conversão em decisões agronômicas (Wolfert *et al.*, 2017). No setor sucroenergético, a etapa de Corte, Transbordo, Transporte e Administração da cana-de-açúcar (CTTA) concentra parte expressiva dos custos operacionais e das emissões de gás carbônico, sobretudo na etapa do transporte rodoviário entre a área agrícola e a unidade industrial. A operação mobiliza comboios de alta capacidade (rodotrens e treminhões). O consumo de combustível corresponde a uma parcela relevante do custo por tonelada transportada e é sensível a fatores como distância, carga, perfil de condução, calibração de pneus e manutenção do motor (Bartholomeu; Péra; Caixeta-Filho, 2016).

A disseminação de sistemas de telemetria embarcada disponibilizou um fluxo contínuo de dados operacionais (quilometragem, consumo, peso transportado, tempos de marcha lenta, frenagem, inércia, entre outros) por viagem e por veículo. Apesar disso, a literatura aponta que o gargalo dominante não é a aquisição do dado, mas, sim, a sua transformação em informação útil para o tomador de decisão (Kamilaris; Kartakoullis; Prenafeta-Boldú, 2017; Tyrychtr; Ulman; Vostrovský, 2015). Em pequenas e médias operações agrícolas, a baixa maturidade analítica e a escassez de profissionais dedicados à análise de dados limitam o aproveitamento do potencial da telemetria (Tyrychtr; Ulman; Vostrovský, 2015). Ferramentas de melhoria contínua, por sua vez, já foram aplicadas com sucesso no monitoramento da colheita mecanizada da cana-de-açúcar, demonstrando que agregam valor à gestão de indicadores agrícolas (Silva *et al.*, 2008; Voltarelli *et al.*, 2018).

2.2 MODELOS DE LINGUAGEM DE GRANDE ESCALA E CONVERSÃO DE LINGUAGEM NATURAL EM LINGUAGEM DE CONSULTA ESTRUTURADA (*TEXT-TO-SQL*)

Os modelos de linguagem de grande escala derivam da arquitetura *Transformer*, que é baseada inteiramente em mecanismos de atenção (Vaswani *et al.*, 2017), e exibem comportamento de aprendizado a partir de poucos exemplos no próprio contexto de entrada (Brown *et al.*, 2020). Sobre essa base, a tarefa de *Text-to-SQL* (traduzir uma pergunta de linguagem natural em uma consulta SQL executável) tornou-se o elo entre o usuário não técnico e os bancos de dados relacionais. Os chamados *benchmarks* de referência estruturaram a avaliação da área: o *Spider* introduziu a avaliação interdomínio com esquemas não vistos no treinamento (Yu *et al.*, 2018); o BIRD aproximou a tarefa de bases de dados reais e de grande escala (Li *et al.*, 2023); e o *Spider* 2.0 evidenciou a dificuldade de fluxos empresariais realistas, nos quais mesmo os modelos no estado da arte apresentam desempenho modesto (Lei *et al.*, 2025).

Em abordagens baseadas em *prompt*, a decomposição da tarefa em subproblemas mostrou-se eficaz: o método DIN-SQL, ao decompor a geração em etapas de *in-context learning* com *self-correction*, elevou a acurácia de execução a um nível próximo ao estado da arte sem ajuste fino do modelo (Pourreza; Rafiei, 2023). Avaliações sistemáticas confirmam que LLMs alcançam acurácia superior a 80% em *benchmarks* abertos, mas degradam em esquemas complexos, assim como em domínios especializados não cobertos no pré-treinamento (Gao, D. *et al.*, 2024). Dois desafios recorrentes dentro desses cenários são a vinculação ao esquema (*schema linking*), isto é, o correto mapeamento dos termos da pergunta para as colunas e tabelas reais, bem como a interpretação semântica da intenção do usuário; ambos são críticos quando o vocabulário operacional do domínio difere da nomenclatura técnica do banco.

2.3 GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG)

A Geração Aumentada por Recuperação combina memória paramétrica, codificada nos pesos do modelo, com memória não paramétrica, acessível por recuperação em um índice vetorial externo (Lewis *et al.*, 2020). Em vez de depender exclusivamente do conhecimento internalizado no treinamento dos grandes modelos de linguagem, a RAG recupera fragmentos relevantes por similaridade vetorial entre *embeddings*, representações numéricas em que textos de significado semelhante ocupam posições próximas, e os injeta no contexto de geração, o que reduz

alucinações, melhora a precisão factual e permite atualizar o conhecimento sem retreinamento (Gao, Y. *et al.*, 2024; Lewis *et al.*, 2020). Aplicada ao *Text-to-SQL*, a RAG recupera exemplos de pares pergunta–consulta previamente validados e os apresenta ao modelo como referência, atuando, desse modo, como um mecanismo de especialização de baixo custo para o domínio-alvo.

A eficácia da RAG depende criticamente da qualidade da recuperação: exemplos de alta similaridade funcionam como modelos diretos para adaptação, ao passo que exemplos de baixa similaridade podem introduzir ruído e degradar a geração de contexto (Gao, Y. *et al.*, 2024). A busca semântica, sustentada por representações vetoriais densas (*embeddings*) e por índices de vizinhança aproximada, é o componente que viabiliza essa recuperação em escala. No contexto operacional, a curadoria de uma base de exemplos validados por especialistas é o que confere à RAG seu valor de especialização e, ao mesmo tempo, é o que determina seu custo de infraestrutura, discutido a seguir.

2.4 CUSTO COMPUTACIONAL, SUSTENTABILIDADE E CONFORMIDADE LGPD

A operação de LLMs em escala desloca o foco do desempenho bruto para a eficiência: o custo por resposta correta tornou-se uma métrica de projeto tão relevante quanto a acurácia (Chen; Zaharia; Zou, 2023). Esse custo decompõe-se em uma parcela variável, proporcional aos *tokens* processados (o *token* é a unidade mínima de texto processada pelo modelo, correspondente a uma palavra ou a um fragmento de palavra), e em uma parcela fixa de infraestrutura, associada a recursos persistentes como *endpoints* de busca vetorial. A proporção entre custo fixo e custo variável determina o volume de consultas necessário para diluir a parcela fixa, de modo que arquiteturas de baixo custo fixo tendem a viabilizar a adoção em operações de menor porte.

A preocupação ambiental acompanha a econômica. A pegada de carbono da inferência, gerada nos centros de dados (*data centers*) que executam os LLMs, pode ser estimada a partir do número de *tokens* processados, da energia consumida por *token*, da eficiência energética da instalação (PUE) e da intensidade de carbono da matriz elétrica local (Patterson *et al.*, 2022). Centros de dados supridos por energia livre de carbono (*Carbon-Free Energy*) reduzem expressivamente as emissões por consulta.

Por fim, a conformidade com a LGPD (Brasil, 2018) é requisito incontornável para sistemas que processam dados operacionais associáveis a operadores. A proteção de dados desde a concepção (*privacy by design*) é assegurada pelas camadas determinísticas de verificação, complementadas pela engenharia de instruções (*prompt engineering*), o que previne a exposição de dados identificáveis sem comprometer a utilidade operacional das respostas.

2.5 CICLO PDCA E GOVERNANÇA DE INTELIGÊNCIA ARTIFICIAL (ISO/IEC 42001)

O ciclo PDCA (Planejar, Fazer, Checar e Agir) é um arcabouço consagrado de melhoria contínua, cuja evolução histórica o consolidou como método de aprendizado iterativo aplicável a processos de naturezas diversas (Moen; Norman, 2009). No agronegócio, sua aplicação ao controle de qualidade de operações mecanizadas é bem documentada (Silva *et al.*, 2008; Voltarelli *et al.*, 2018). Transposto para a arquitetura de um sistema de inteligência artificial, o ciclo oferece uma decomposição natural da tarefa em fases auditáveis: planejar a interpretação da intenção, fazer a geração da consulta, checar a validade e a aderência dos dados retornados e agir na entrega da resposta ao usuário.

Essa decomposição dialoga diretamente com a norma ISO/IEC 42001, primeira norma internacional de governança de sistema de gestão para inteligência artificial, que enfatiza rastreabilidade, melhoria contínua e responsabilização ao longo do ciclo de vida do sistema (Thiers; Harned, 2024). Ao mapear cada agente a uma fase do PDCA, o sistema incorpora, por construção, pontos de verificação e registro que sustentam a governança auditável preconizada pela norma. É nessa convergência entre um método clássico de melhoria contínua e um referencial contemporâneo de governança de inteligência artificial que se assenta a proposta metodológica deste trabalho, detalhada a seguir.

3 MATERIAL E MÉTODOS

3.1 FONTE E NATUREZA DOS DADOS

Os dados foram obtidos por meio de sistemas de telemetria embarcada em caminhões durante as operações de transporte de colmos de cana-de-açúcar, da área agrícola à agroindústria em uma unidade sucroenergética parceira. Após a chegada do rodotrem ao centro de descarregamento industrial dos colmos de cana-de-açúcar, os caminhões passam por pesagem e transmitem os dados via rede 4G. As variáveis operacionais coletadas por viagem foram: quilometragem percorrida, volume de combustível consumido, peso bruto transportado, peso tara do caminhão, modelo do caminhão, identificação do caminhão e do motorista, bloco e turno da operação logística, e atributos da zona de manejo (fazenda, zona e talhão), bem como indicadores históricos de telemetria por viagem (tempos relativos de inércia, frenagem e marcha lenta, além da carga do motor e dos tempos de operação dentro ou acima da faixa econômica). A coleta ocorreu ao longo da safra de 2025; o experimento de avaliação do sistema, descrito adiante, foi, por sua vez, conduzido em abril de 2026 sobre esses dados já consolidados.

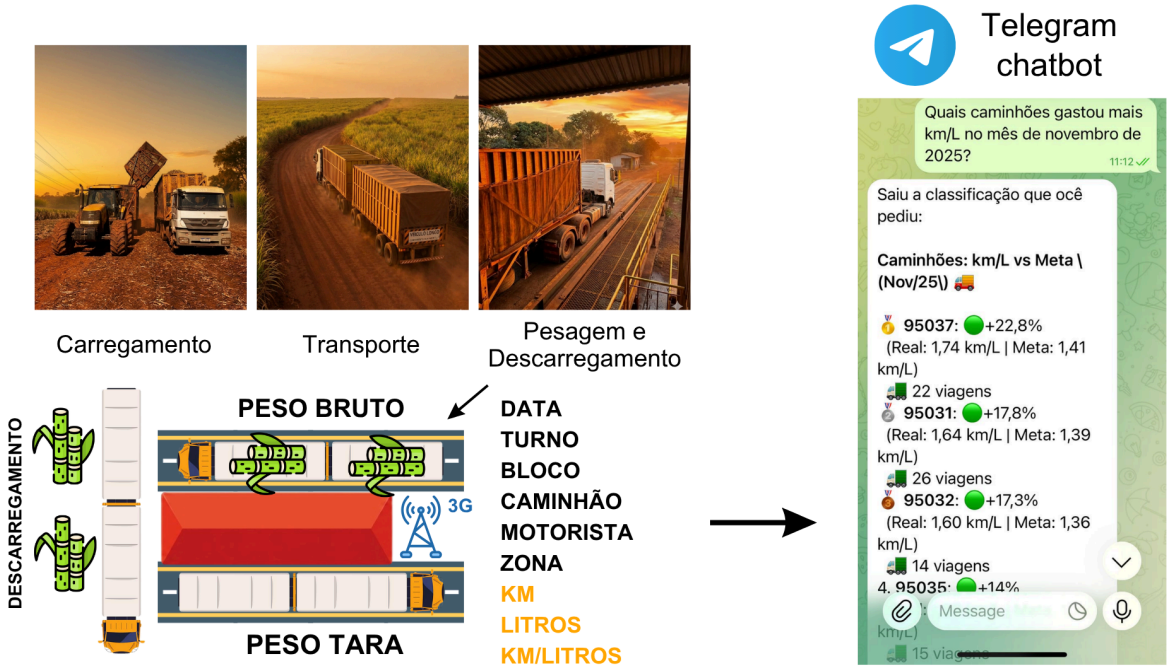
Após a etapa de limpeza dos dados brutos, foram retidas 30.791 viagens válidas. O tratamento de dados anômalos adotou uma regra determinística de exclusão de *outliers* baseada na coerência física entre distância e consumo: viagens cuja quilometragem registrada divergia em mais de 30% da distância de referência da rota foram sinalizadas e removidas, visto que indicavam falha de registro do odômetro ou desvios não representativos da operação regular.

Esse limiar foi escolhido por equilibrar a preservação da variabilidade legítima da operação, inerente a rotas e condições de campo distintas, com a remoção de registros fisicamente inconsistentes, evitando que medições espúrias contaminassem tanto as estatísticas descritivas quanto o modelo preditivo. A camada de exclusão foi aplicada de forma transparente e auditável, integrando o filtro de qualidade como primeira condição de toda consulta sobre a base operacional.

A base operacional completa, anterior à exclusão de valores discrepantes, reúne 43.112 viagens de transporte de cana-de-açúcar, com peso líquido médio de 59,64 toneladas por viagem, totalizando 2.571.282 toneladas movimentadas no

período. Aplicando o custo de referência de R$ 4,10 por tonelada transportada (Pecege, 2026), o dispêndio operacional total é da ordem de R$ 10,5 milhões, valor adotado como referência econômica para a análise de custo da seção 4.2.

Figura 1 – Fluxo operacional do transporte de cana-de-açúcar, da área agrícola à agroindústria, e interface conversacional do sistema multiagente via mensageria *Telegram*

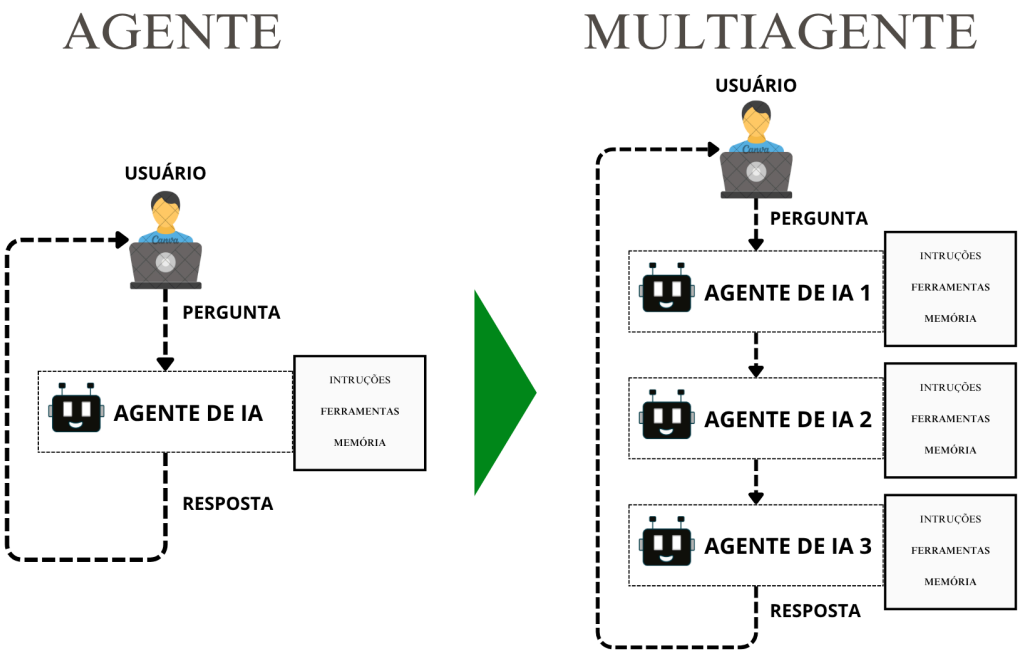


Fonte: Elaborada pelo autor (2026).

3.2 ARQUITETURA DO SISTEMA MULTIAGENTE

O sistema, denominado agente PDCA, foi implementado como aplicação conteinerizada em ambiente de nuvem, com processamento assíncrono e orquestração de fluxo de automação entre quatro agentes especializados. A opção pela arquitetura multiagente decorre da natureza composta da tarefa: em vez de delegar a uma única chamada de modelo a interpretação da pergunta, a geração da consulta, a verificação dos dados e a formatação da resposta, a decomposição distribui essas responsabilidades entre agentes dedicados, cada qual otimizável e auditável de forma independente (Guo *et al.*, 2024). O contraste conceitual entre as duas abordagens é sintetizado na Figura 2.

Figura 2 – Comparação conceitual entre a arquitetura de agente único e a arquitetura multiagente, evidenciando a decomposição da tarefa complexa em agentes especializados



Fonte: Adaptada de Guo *et al.* (2024).

Os quatro agentes foram mapeados às fases do ciclo PDCA, conferindo ao sistema a estrutura de melhoria contínua e os pontos de verificação preconizados pela norma ISO/IEC 42001 (Thiers; Harned, 2024). O agente Planejar (*Plan*) recebe a pergunta do usuário sobre o consumo dos caminhões, classifica a intenção do usuário e estrutura um plano analítico; o agente Fazer (*Do*) gera a consulta SQL, eventualmente apoiado por exemplos recuperados via RAG, e a submete a uma verificação prévia de validade da consulta SQL (*dry-run*); o agente Checar (*Check*) executa a consulta, valida a aderência dos dados retornados à pergunta e formata o relatório; e o agente Agir (*Act*) realiza a verificação final de conformidade com a LGPD e entrega a resposta em relação ao consumo de combustível dos caminhões ao usuário. Esse mapeamento e o respectivo alinhamento normativo são detalhados na Figura 3.

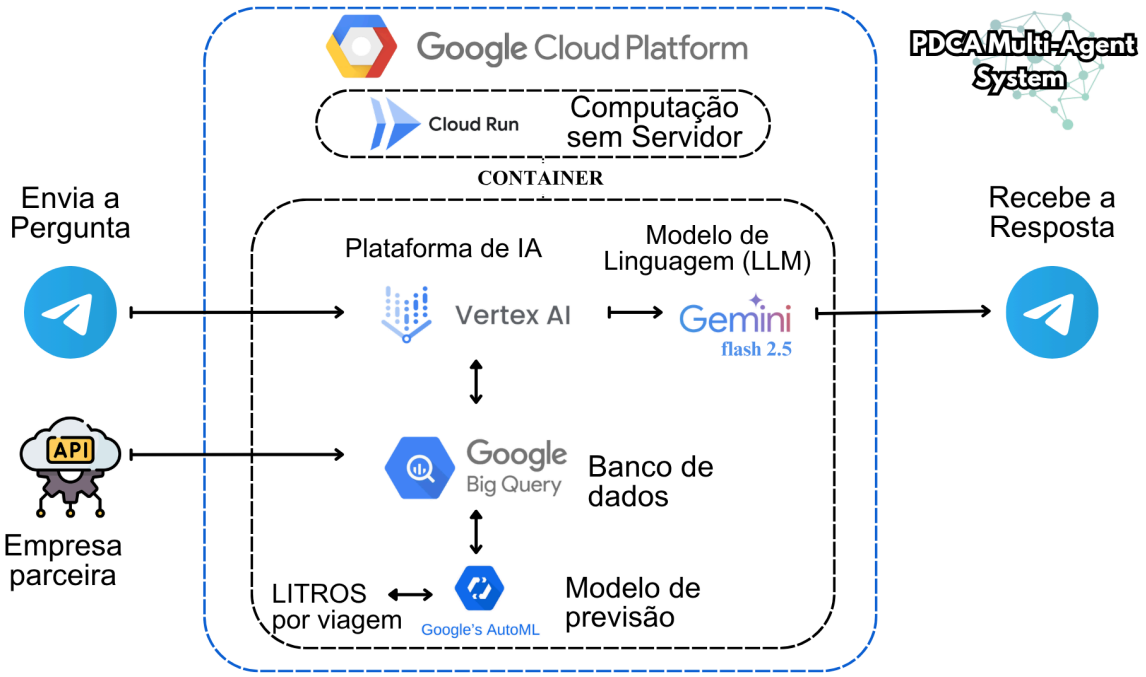
Figura 3 – Mapeamento dos quatro agentes especializados às fases do ciclo PDCA (Planejar, Fazer, Checar e Agir), com os respectivos papéis e o alinhamento à norma ISO/IEC 42001

Fonte: Elaborada pelo autor (2026).

Toda a infraestrutura baseia-se no modelo *serverless* e é executada integralmente na nuvem *Google Cloud Platform* (GCP), sem dependência de *hardware* local. Os agentes foram implantados como serviços conteinerizados no *Cloud Run*, e o processamento das mensagens é desacoplado por meio de uma fila assíncrona (*Cloud Tasks*), de modo que o recebimento da pergunta e a execução do ciclo PDCA ocorrem em componentes distintos. Essa abordagem garante resposta imediata ao usuário e escalabilidade elástica sob carga variável, ou seja, a capacidade de o sistema provisionar e liberar automaticamente instâncias de contêiner conforme o volume de requisições oscila, ampliando os recursos nos picos de consulta durante a safra e reduzindo-os até zero nos períodos ociosos, sem reconfiguração manual nem custo por capacidade inativa.

O modelo de linguagem empregado em todos os agentes generativos foi o *Gemini 2.5 Flash*, acessado via plataforma de inteligência artificial generativa *Vertex AI*; o armazenamento e a consulta dos dados operacionais foram realizados em armazém de dados analítico (*BigQuery*), que também hospeda o modelo preditivo de consumo treinado por aprendizado de máquina automatizado (*AutoML do BigQuery*, detalhado na seção 3.3); e a recuperação de exemplos validados apoiou-se em um índice de busca vetorial (*Vertex AI Vector Search*). O fluxo completo de uma consulta, do recebimento via *Telegram* ao envio da resposta, percorre os estágios encadeados representados na Figura 4.

Figura 4 – Arquitetura *serverless* do sistema multiagente na plataforma de nuvem *Google Cloud Platform*



Fonte: Elaborada pelo autor (2026).

3.3 MODELO DE PREDIÇÃO DO CONSUMO DE COMBUSTÍVEL

O modelo preditivo foi treinado com o *AutoML* do *BigQuery* (*BigQuery ML*), recurso que automatiza a construção do modelo a partir de dados estruturados: avalia diferentes famílias de algoritmos, como *ensembles* de árvores de decisão com *gradient boosting* e redes neurais artificiais (*deep neural networks*), e ajusta automaticamente arquiteturas e hiperparâmetros. Esse ajuste automático de hiperparâmetros, embutido no *AutoML Tables* por padrão, aciona uma divisão tripla dos dados, que são embaralhados aleatoriamente e repartidos em 80% para treino, 10% para validação e 10% para teste (*Google Cloud*, 2026c). O treinamento ajusta os modelos candidatos, a validação orienta a seleção da melhor configuração de modelo e hiperparâmetros, e o teste, reservado e não visto durante o ajuste, fornece as métricas finais de desempenho.

A configuração vencedora alcançou coeficiente de determinação R^2 = 0,98, com erro médio absoluto (MAE) de 1,63 L e raiz do erro quadrático médio (RMSE) de 2,24 L, indicando elevada capacidade explicativa sobre o consumo por viagem.

A variável-resposta é o consumo de combustível por viagem. Entre os preditores, as variáveis categóricas (modelo do caminhão e zona de operação) exigem conversão para uma representação matemática adequada aos algoritmos de aprendizado. Essa transformação, denominada codificação *one-hot* (ou criação de variáveis *dummy*), converte cada categoria em vetores binários: para a variável modelo do caminhão, por exemplo, cria-se uma coluna para cada modelo existente, atribuindo-se o valor 1 quando o transporte foi realizado por aquele veículo específico e 0 nos demais casos. Visto que cada variável original se desdobra em múltiplos componentes binários, a Tabela 1 apresenta a atribuição de pesos reagrupada por variável de origem, e não por categoria isolada.

O peso de cada variável nas previsões foi medido pelo recurso de inteligência artificial explicável (*Explainable AI*) do *BigQuery*, que quantifica o quanto cada informação de entrada contribuiu para o consumo estimado em cada viagem. Essa contribuição é chamada de atribuição, e a importância global de uma variável corresponde à média dessas atribuições, em valor absoluto, ao longo de todas as viagens avaliadas: quanto maior esse valor, maior a influência da variável sobre o modelo. O cálculo baseia-se nos valores de *Shapley* (Shapley, 1953), conceito originário da teoria dos jogos, o qual reparte de forma equilibrada o resultado previsto entre as variáveis, segundo a contribuição média de cada uma. Por se tratar de um modelo com muitas variáveis, essa repartição é estimada por amostragem (Amostragem de *Shapley*), procedimento adotado pela plataforma (*Google Cloud*, 2026a) que adapta os valores de Shapley à atribuição de importância em aprendizado de máquina (Lundberg; Lee, 2017). A Tabela 1 apresenta o resultado dessa análise em valor de atribuição absoluto e em importância relativa, isto é, a proporção que cada variável representa do total.

Conforme a Tabela 1, a quilometragem da viagem destacou-se como variável dominante, respondendo por 60,4% da importância total do modelo. Sua atribuição média absoluta foi de 12,18 L, cerca de seis vezes a do segundo preditor mais relevante. Em um segundo patamar, próximos entre si, situaram-se o tempo histórico em inércia (9,6%) e o tempo histórico de operação dentro da faixa econômica (9,3%). Seguiram-se, em ordem decrescente, a carga média histórica do motor (4,9%) e o modelo do caminhão (3,6%), ao passo que os demais preditores (tempo acima da faixa econômica, zona de operação, paradas, marcha lenta e frenagem) apresentaram contribuições individuais inferiores a 3%.

Tabela 1 – Importância das variáveis preditoras do modelo de consumo de combustível (*AutoML* do *BigQuery*), obtida por inteligência artificial explicável (*Explainable AI*), expressa em valor de atribuição absoluto e em importância relativa (proporção da atribuição total)

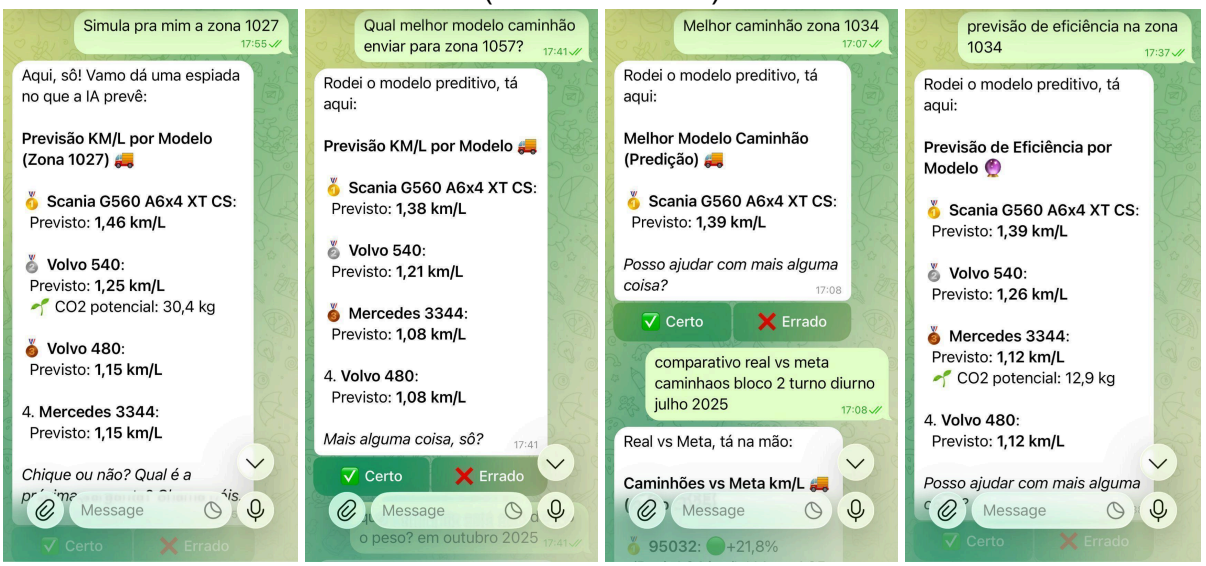
Variável preditora (recurso)	Atribuição (L)	Importância relativa (%)
Quilometragem da viagem (KM)	12,181	60,4
Tempo histórico em inércia (MEDIA_HIST_INERCIA)	1,934	9,6
Tempo histórico dentro da faixa econômica (MEDIA_HIST_DENTRO_ECON)	1,875	9,3
Carga média histórica do motor (MEDIA_HIST_CARGA_MOTOR)	0,987	4,9
Modelo do caminhão (MODELO_CAMINHAO)	0,718	3,6
Tempo histórico acima da faixa econômica (MEDIA_HIST_ACIMA_ECON)	0,578	2,9
Zona de operação (ZONA)	0,561	2,8
Tempo histórico em paradas (MEDIA_HIST_PARADAS)	0,541	2,7
Tempo histórico em marcha lenta (MEDIA_HIST_MARCHA_LENTA)	0,412	2,0
Tempo histórico em frenagem (MEDIA_HIST_FRENAGEM)	0,378	1,9
Total	**20,165**	**100,0**

Fonte: Elaborada pelo autor (2026).

Em conjunto, esses resultados confirmam que o modelo captura de forma consistente os determinantes do consumo de combustível no transporte canavieiro: a quilometragem responde pela maior parte da variação absoluta, ao passo que os indicadores de perfil de condução (tempo em inércia, operação dentro da faixa econômica e carga do motor) atuam como moduladores da eficiência, justamente as variáveis passíveis de intervenção operacional.

Essa distinção é o que confere valor prático ao modelo, pois permite separar o consumo estruturalmente imposto pela rota daquele que pode ser reduzido por ajustes de condução e de alocação da frota. Integrado ao agente PDCA, o modelo traduz essas estimativas em cenários de eficiência por zona de operação e por modelo de caminhão, disponibilizados ao usuário diretamente na interface conversacional (Figura 5), de modo a subsidiar decisões que reduzam o consumo de diesel e, consequentemente, os custos operacionais e as emissões de CO_2 associadas ao transporte da cana-de-açúcar.

Figura 5 – Exemplos de resposta do agente PDCA na interface do *Telegram*, com previsão de eficiência por modelo de caminhão e o mecanismo de validação humana (certo ou errado)



Fonte: Elaborada pelo autor (2026).

3.4 DELINEAMENTO EXPERIMENTAL

A avaliação seguiu um delineamento fatorial 2 × 2, cruzando dois fatores. O primeiro fator consistiu na estratégia de engenharia de *prompt*, avaliada em dois níveis: inferência sem exemplos (*zero-shot*) e abordagem com exemplos contextuais (*few-shot*). Esta última modalidade foi operacionalizada por meio da Geração Aumentada por Recuperação (*Retrieval-Augmented Generation*, RAG), a partir de uma base de 90 pares pergunta–consulta previamente validados por um especialista e selecionados via similaridade vetorial.

O segundo fator foi a composição do conjunto de perguntas, também em dois níveis. O primeiro nível compreendeu um conjunto de 900 perguntas, derivado da variação sistemática de uma lista de perguntas operacionais elaborada pelo supervisor da frota de caminhões da empresa parceira em conjunto com os instrutores dos motoristas dos caminhões canavieiros, profissionais cuja função é treinar os condutores com o objetivo de alcançar viagens cada vez mais econômicas e seguras, ou seja, mais sustentáveis para a operação agrícola. Essa lista, fornecida pela empresa parceira e representativa das demandas do dia a dia da operação, foi expandida por meio de combinações reais extraídas do armazém de dados (pares motorista–zona, trios caminhão–bloco–turno e demais combinações operacionais), a fim de cobrir a diversidade de recortes efetivamente consultados em campo. O

segundo nível consistiu em um subconjunto de 732 perguntas, extraído por meio da exclusão dos casos que obtiveram resultado legitimamente vazio, isto é, perguntas cujo recorte não retornaria dados, visto que a combinação não existia. O cruzamento dos dois fatores resultou em quatro condições experimentais (ZS-900, FS-900, ZS-732 e FS-732), totalizando 3.264 execuções automatizadas via máquina virtual conduzidas entre os dias 18 e 21 de abril de 2026, no decorrer de quatro testes que somaram 36,2 horas de operação ininterrupta do sistema.

A instrumentação do sistema foi concebida para que cada interação processada gerasse um registro completo e autocontido, de modo que a totalidade do experimento pudesse ser reconstruída e auditada exclusivamente a partir dos dados armazenados. Para isso, toda consulta submetida ao agente PDCA foi gravada em uma tabela de registros de interação (*interaction logs*) no *BigQuery*, na qual cada linha corresponde a uma execução e consolida, além dos campos primários, um conjunto de metadados aninhados que descrevem as etapas do ciclo de processamento de cada pergunta e resposta.

Os campos primários compreendem três blocos. O primeiro trata da identificação e da rastreabilidade da interação: identificador único, carimbo de tempo (*timestamp*), versão do sistema e dados de roteamento do *Telegram*, como número do usuário e número da mensagem. O segundo registra a entrada e o processamento linguístico: a pergunta original do usuário, a pergunta reescrita pelo agente e a intenção estruturada em dimensões, filtros, indicadores e período. O terceiro documenta a geração e a execução da consulta: o SQL gerado, a cadeia de raciocínio (*chain of thought*) associada, a configuração de visualização, a resposta final enviada ao usuário, o número de tentativas (*retry*) e os sinalizadores de erro e de truncamento de dados por motivos de conformidade com a LGPD.

Os metadados aninhados, por sua vez, constituem a base quantitativa das análises conduzidas posteriormente e abrangem quatro dimensões. A primeira registra o consumo de *tokens* discriminado por agente do ciclo PDCA (*Plan*, *Do*, *Check* e *Act*), distinguindo *tokens* de entrada (*input*), de saída (*output*), de raciocínio (*thinking*) e em cache (*cached*), o que fundamenta a estimativa de custo. A segunda contabiliza a latência total e por etapa, incluindo o tempo de recuperação vetorial e o de registro final. A terceira documenta os custos de processamento no *BigQuery*, com os *bytes* faturados e a ocorrência de acerto de cache. A quarta detalha a Geração Aumentada por Recuperação (RAG): a estratégia adotada (*zero-shot* ou

few-shot por *template*), o número de exemplos utilizados, a menor distância obtida, os identificadores dos exemplos recuperados e os respectivos vetores de *embedding*. Complementarmente, uma tabela acessória de recuperação registra, a cada consulta ao índice vetorial, o texto consultado, o número de vizinhos solicitados (*top-k*), a métrica de similaridade por cosseno e os exemplos retornados com as respectivas distâncias.

A execução automatizada das perguntas e a coleta da verdade de referência (*ground truth*) foram conduzidas de forma programática a partir da base operacional, garantindo que cada uma das 3.264 execuções fosse acompanhada de seus metadados completos. Esse histórico estruturado constitui o registro único a partir do qual foram derivadas todas as métricas de acurácia, custo, latência, pegada de carbono e impacto da RAG apresentadas posteriormente.

A validação do estimador de custo apoiou-se no confronto entre duas fontes independentes, construídas por caminhos opostos. O custo variável estimado foi obtido de forma ascendente (*bottom-up*), a partir dos metadados de consumo registrados em cada uma das 3.264 execuções: para o modelo de linguagem, os *tokens* de entrada, de saída, de raciocínio (*thinking*) e em cache de cada agente do ciclo PDCA foram multiplicados pelos respectivos preços unitários publicados; os demais serviços (busca vetorial e consultas analíticas) foram contabilizados de modo análogo, a partir de suas unidades de cobrança. Todos os valores foram convertidos para reais pela taxa de câmbio de faturamento aplicada pelo *Google Cloud* no período do experimento (R$ 5,9522/US$), registrada no campo *currency_conversion_rate* do arquivo de exportação de faturamento para o *BigQuery* e idêntica à taxa que converteu a fatura real validada neste trabalho (*Google Cloud*, 2026b). Essa escolha garante a coerência entre o custo estimado e o custo real, ambos na mesma base cambial. O custo variável real, por sua vez, foi obtido de forma descendente (*top-down*), diretamente do faturamento do *Google Cloud Platform*: no módulo financeiro do console (*Billing*), exportou-se o relatório detalhado de custos discriminado por SKU (*stock keeping unit*), restrito ao mesmo período e aos mesmos serviços do experimento.

Como camada adicional de confiabilidade e auditabilidade, todas as execuções passaram por validação humana conduzida pelo autor a partir da tabela de registros de atividade (*interaction logs*). Para cada registro, conferiu-se manualmente a correspondência entre a pergunta formulada, a consulta SQL gerada

e a resposta entregue ao usuário, o que assegurou que o acerto contabilizado refletisse uma resposta efetivamente correta, e não apenas uma consulta executável.

3.4.1 Ambiente de execução automatizada e orquestração do experimento

A condução das 3.264 execuções de forma programática e reprodutível exigiu a separação física entre a automação de execução das perguntas e o sistema sob avaliação (agente PDCA). Essa separação é metodologicamente relevante: visto que o objetivo incluía aferir o custo e a latência do sistema multiagente em condições equivalentes às de produção (seções 3.5 e 4.2), o componente que dispara e contabiliza o experimento não pode residir na mesma infraestrutura medida, sob pena de contaminar as métricas com o processamento do próprio teste. Adotou-se, portanto, uma arquitetura em que o orquestrador é permanente e externo, ao passo que o sistema avaliado permanece integralmente sob demanda (*serverless*).

O agente denominado *quest-runner* foi implementado em uma máquina virtual dedicada do serviço de computação em nuvem (*Compute Engine*), provisionada na zona *us-central1-a* da mesma região do restante da infraestrutura, com encaminhamento de endereço IP desabilitado e autenticação por conta de serviço gerenciada. Diferentemente dos agentes do ciclo PDCA, que são acionados sob demanda e liberados após cada consulta, a máquina virtual constitui um recurso de disponibilidade contínua, condição necessária para sustentar testes de longa duração, de até 571 minutos por bateria de 900 perguntas, sem interrupção e para preservar o estado de progresso entre as execuções.

O fluxo de avaliação operou em três etapas encadeadas. Na primeira, o agente *quest-runner* recuperava as perguntas de uma tabela dedicada do armazém de dados (*BigQuery*), fisicamente separada das tabelas operacionais e de registros de atividade do sistema, na qual cada linha correspondia a uma pergunta do conjunto experimental, acompanhada de seu identificador e dos metadados de estratificação (condição, tipo e nível de complexidade descritos na seção 3.4.3). Essa tabela funcionou como fila canônica do experimento, o que garantiu que cada uma das 3.264 execuções fosse rastreável a um item de origem e que as quatro

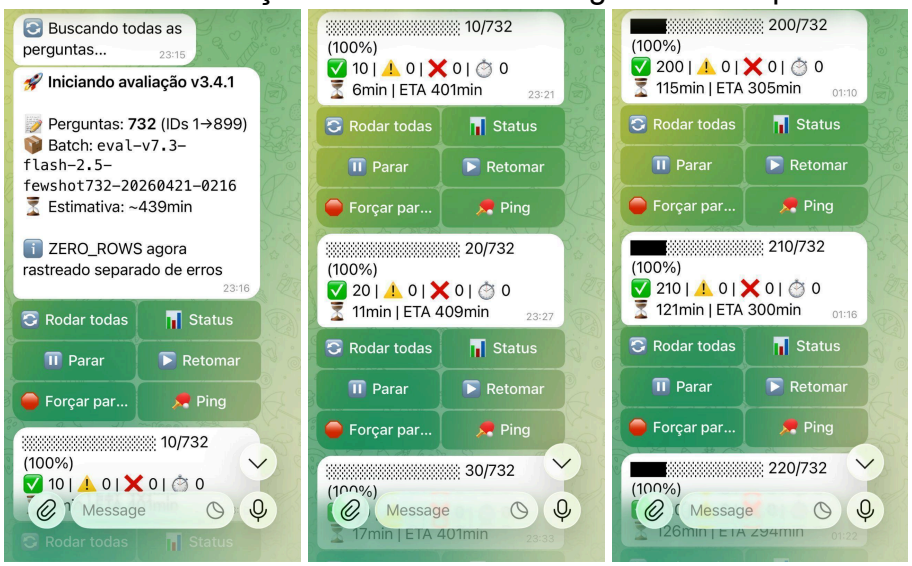
condições (ZS-900, FS-900, ZS-732 e FS-732) pudessem ser disparadas, pausadas e retomadas de forma controlada.

Na segunda etapa, o agente *quest-runner* submetia cada pergunta ao sistema multiagente por meio da própria interface de mensageria (*Telegram*), reproduzindo fielmente o canal de uso real do supervisor de frota, sem qualquer atalho interno que contornasse o ciclo completo. Para isso, o agente atuava como um cliente de mensageria programática, emulando o envio de mensagens de um usuário humano e capturando as respostas devolvidas pelo sistema. Esse desenho assegura validade externa às métricas, uma vez que a pergunta percorre exatamente o mesmo caminho de uma consulta de produção: do recebimento pelo *webhook* ao enfileiramento assíncrono e à execução do ciclo PDCA.

Na terceira etapa, o agente PDCA, hospedado como serviço conteinerizado no *Cloud Run* (seção 3.2), recebia a pergunta, processava integralmente as fases Planejar, Fazer, Checar e Agir e devolvia a resposta ao *Telegram*, ao mesmo tempo em que registrava o conjunto completo de metadados da execução na tabela de registros de atividade (*interaction logs*). Enquanto isso, o agente *quest-runner* monitorava o retorno de cada resposta, controlava o ritmo de submissão visando respeitar os limites de concorrência da fila assíncrona e consolidava o avanço da bateria, expondo indicadores de acompanhamento em tempo real (total processado, acertos, alertas, erros e perguntas com resultado vazio). A Figura 6 ilustra a interface de controle do agente orquestrador, com a identificação da bateria, o conjunto de perguntas em execução, a estimativa de duração e os comandos de gestão do experimento (executar, pausar, retomar, interromper e verificar disponibilidade).

Por fim, a validação humana de cada resposta, conduzida pelo autor a partir dos registros (seção 3.4), era reconciliada com a tabela de perguntas de origem: o veredito de correção, sinalizado na própria interface de mensageria, era propagado de volta ao item correspondente da fila experimental, o que fechava o ciclo entre pergunta planejada, execução registrada e julgamento de correção. Essa arquitetura de orquestração externa, ancorada na máquina virtual *quest-runner*, foi o que conferiu ao experimento a escala, a reprodutibilidade e a rastreabilidade necessárias à derivação das métricas apresentadas na seção 4.

Figura 6 – Interface de controle do agente *quest-runner* no *Telegram*, demonstrando a identificação da bateria de avaliação, o conjunto de perguntas em execução, a estimativa de duração e os comandos de gestão do experimento



Fonte: Elaborada pelo autor (2026).

3.4.2 Base de exemplos validados e recuperação vetorial

A base de exemplos validados constitui o insumo que confere especialização à Geração Aumentada por Recuperação (RAG). Foram reunidos 90 pares de pergunta e consulta SQL, conferidos manualmente um a um pelo autor e armazenados na tabela *fewshot_examples* do armazém de dados (BigQuery). Cada par recebeu um identificador próprio (*validation_id*), que garante a rastreabilidade de qual exemplo foi utilizado em cada recuperação. Para que o sistema compare as perguntas pelo significado, e não apenas por palavras idênticas, o texto de cada exemplo é convertido em uma representação numérica (*embedding*), isto é, uma lista de 768 números que posiciona a frase em um espaço no qual frases de sentido análogo mantêm-se próximas entre si.

Esses vetores são guardados em um banco de dados vetorial (*Vertex AI Vector Search*). Visto que o conjunto é reduzido, com apenas 90 exemplos, adotou-se um índice por força bruta (*brute force*), que compara a pergunta recebida com todos os registros, sem aproximações, o que garante a recuperação exata dos mais semelhantes. A semelhança entre os dois textos é medida pelo produto escalar de vetores previamente normalizados (normalização L2), procedimento equivalente à similaridade por cosseno, na qual o valor resultante é tanto maior quanto menor o ângulo entre os vetores. A cada pergunta, as sentenças mais próximas (vizinhos

mais próximos) são recuperadas e injetadas como referência no agente Fazer, orientando a geração da consulta SQL. A base é atualizada em tempo real (*streaming*), de modo que novos exemplos validados tornam-se consultáveis imediatamente. Por exigir um ponto de acesso (*endpoint*) dedicado e continuamente disponível, esse mecanismo constitui o principal componente de custo fixo do sistema.

3.4.3 Estratificação por complexidade da consulta SQL

A distribuição do conjunto de 900 perguntas foi estratificada de forma ortogonal aos dois fatores experimentais, segundo a complexidade da consulta SQL necessária para respondê-las, em três níveis (simples, detalhada e complexa), seguindo uma distribuição balanceada de 300 perguntas por nível (50 perguntas para cada um dos seis tipos de consulta: *ranking,* contagem, soma, média, comparativo e predição). A estratificação teve por finalidade avaliar se a complexidade estrutural da consulta afeta a acurácia, a latência e o custo do sistema, e seguiu os critérios de dificuldade propostos por Yu *et al.* (2018) para a avaliação de sistemas de conversão de linguagem natural em SQL (*Text-to-SQL*), segundo os quais o grau de dificuldade é determinado pelo número e pela natureza dos componentes SQL exigidos.

O nível simples reúne consultas resolvidas por uma única agregação sobre a base operacional, com uma única dimensão de agrupamento e, no máximo, um filtro temporal, além do filtro determinístico de *outlier* comum a todas as consultas. Não há filtros compostos nem cálculos derivados, de modo que a tradução da pergunta para SQL é direta e com baixo risco de ambiguidade. Exemplos representativos desse nível, acompanhados das respectivas respostas do agente PDCA na interface do *Telegram*, são apresentados no Apêndice A (Figura 1A).

O nível detalhado abrange consultas que combinam de duas a três condições de filtro simultâneas (tipicamente cruzamentos de zona, turno, bloco e período), as quais podem exigir janelas temporais compostas (cláusula *BETWEEN*), uma segunda dimensão de agrupamento e a interpretação de termos coloquiais próprios da operação. A geração do SQL torna-se mais sensível à vinculação correta dos filtros ao esquema (*schema linking*), o que eleva o grau de dificuldade em

relação ao nível simples. Exemplos representativos são apresentados no Apêndice A (Figura 2A).

O nível complexo compreende consultas que demandam cálculos derivados (como litros por tonelada transportada), comparação com metas operacionais, cruzamento de múltiplas dimensões (zona, bloco e talhão), subconsultas, expressões condicionais, operações de conjunto, ou o acionamento do modelo preditivo de consumo para a construção de cenários contrafactuais. São as consultas de maior exigência sintática e semântica, nas quais se concentra o maior risco de erro de lógica SQL e de interpretação. Exemplos representativos são apresentados no Apêndice A (Figura 3A).

3.5 MÉTRICAS E ANÁLISE ESTATÍSTICA

As métricas foram organizadas em cinco eixos: (i) acurácia de execução, calculada pela correspondência entre o resultado retornado e o resultado esperado (Li *et al.*, 2023); (ii) categorização de erros, mediante a taxonomia tripartida em vinculação ao esquema (*schema linking*), lógica SQL e interpretação semântica (Lei *et al.*, 2025; Yu *et al.*, 2018); (iii) custo operacional, decomposto em custos fixo e variável; (iv) latência por agente e por nível de complexidade da consulta SQL no banco de dados; e (v) pegada de carbono da inferência.

A análise estatística foi conduzida em *Python*, no ambiente *Google Colab*, com as bibliotecas *SciPy* (Virtanen *et al.*, 2020) e *statsmodels* (Seabold; Perktold, 2010), adotando-se o nível de significância $\alpha = 0,05$. Cada teste cumpriu uma função específica: o teste de *McNemar* exato (Agresti, 2002) comparou, no mesmo conjunto de perguntas, se uma condição acertava onde a outra errava, sendo apropriado para dados pareados de acerto e erro, aplicado à comparação da acurácia entre a geração sem exemplos (*zero-shot*) e a geração com exemplos recuperados (*few-shot*); por sua vez, a associação entre o nível de complexidade da consulta e a ocorrência de erro foi avaliada pelo teste exato de Fisher, na extensão de Freeman e Halton (1951) para tabelas 3×2, calculado por enumeração completa de todas as tabelas com as mesmas margens. Adotou-se o teste exato em lugar do teste de qui-quadrado porque o número de erros por condição foi baixo, de 3 a 9, e as frequências esperadas mínimas ficaram entre 0,89 e 3,00, abaixo do valor 5 a partir do qual a aproximação assintótica do qui-quadrado é considerada confiável (Agresti,

2002). Como medida de tamanho de efeito, reportou-se o *V* de Cramér, interpretado segundo Cohen (1988), para quem o valor 0,10 corresponde a efeito pequeno; valores inferiores foram tratados como efeito desprezível.

O método de Wilson forneceu intervalos de confiança de 95% adequados a proporções próximas de 100%, situação em que os intervalos tradicionais perdem precisão. O teste de *Mann-Whitney U* contrastou a latência entre duas condições, ao passo que o teste de *Kruskal-Wallis* estendeu esse contraste aos três níveis de complexidade da consulta (Hollander; Wolfe; Chicken, 2014). A escolha de testes não paramétricos justifica-se pela natureza discreta da variável acurácia (*feedback* binário de acerto ou erro) e pela distribuição não normal das variáveis contínuas de custo e latência.

A pegada de carbono por consulta foi estimada pela Equação 1 (Patterson *et al.*, 2022):

$$CF_q = T_q \times E_t \times PUE \times CIF \tag{1}$$

em que CF_q é a pegada de carbono por consulta (g CO_2); T_q, o número de *tokens* processados na consulta; E_t, a energia consumida por *token* (kWh *token*$^{-1}$); *PUE*, a eficiência no uso de energia do centro de dados; e *CIF*, a intensidade de carbono da matriz elétrica (g CO_2 kWh^{-1}). Estimaram-se dois cenários: a matriz elétrica bruta e a matriz ajustada por energia livre de carbono (*Carbon-Free Energy*). Os dois cenários diferem apenas no valor de *CIF*: no cenário bruto, *CIF* corresponde à intensidade de carbono média da rede elétrica local, considerando todas as fontes de geração; no cenário ajustado, desconta-se do *CIF* a parcela de fontes renováveis e nuclear que o centro de dados efetivamente consome, o que resulta em uma intensidade de carbono menor.

4 RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados das 3.264 execuções automatizadas de perguntas e respostas conduzidas com base em dados reais de telemetria das 30.791 viagens válidas, avaliados nos eixos de acurácia, custo, latência, desempenho da Geração Aumentada por Recuperação (RAG), pegada de carbono e a governança de inteligência artificial (ISO/IEC 42001) e a proteção de dados pessoais (LGPD).

4.1 ACURÁCIA DE EXECUÇÃO E IMPACTO DA RAG

A acurácia de execução foi igual ou superior a 99,00% em todas as quatro condições experimentais, variando de 99,00% (ZS-900) a 99,59% (FS-732), com intervalos de confiança de 95% pelo método de Wilson reportados na Tabela 2. O teste de McNemar exato pareado não detectou ganho estatisticamente significativo da Geração Aumentada por Recuperação sobre a geração sem exemplos, nem no conjunto de 900 perguntas (discordâncias $b = 6$, $c = 7$; $p = 1{,}00$) nem no de 732 perguntas ($b = 2$, $c = 5$; $p = 0{,}45$).

Tabela 2 – Acurácia de execução por condição experimental, com intervalos de confiança de 95% pelo método de Wilson

Condição	n	Acertos	Acurácia (%)	IC 95% (Wilson)
ZS-900	900	891	99,00	[98,1–99,5]
FS-900	900	892	99,11	[98,3–99,5]
ZS-732	732	726	99,18	[98,2–99,6]
FS-732	732	729	99,59	[98,8–99,9]

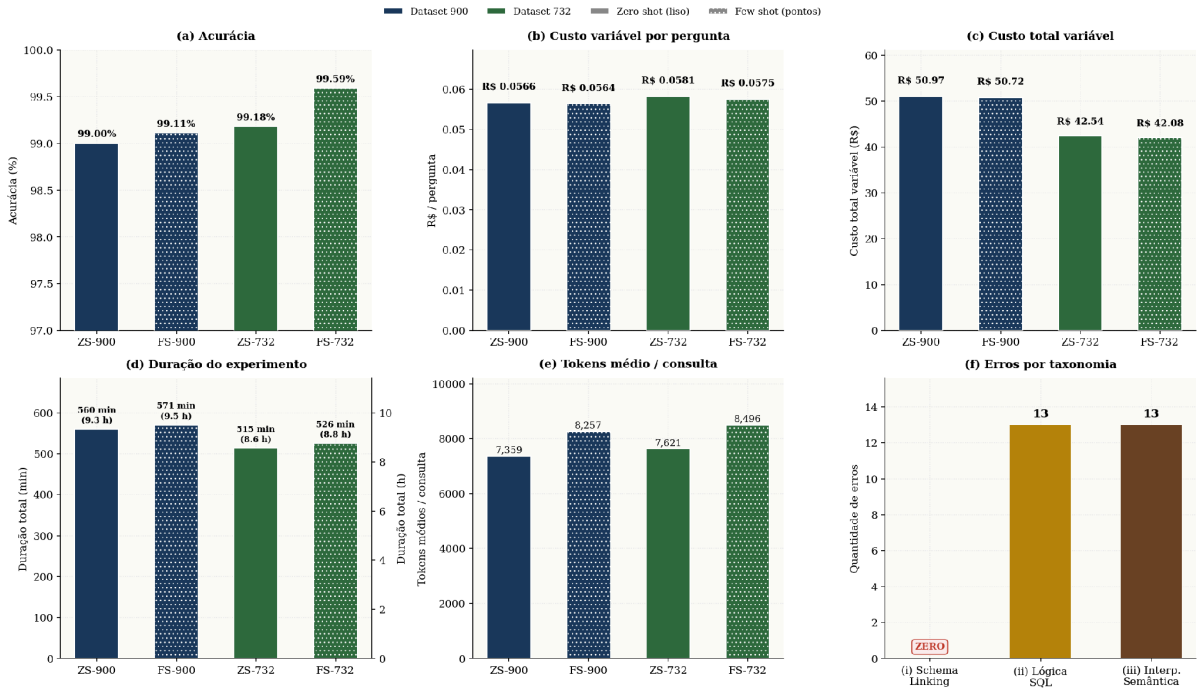
Fonte: Elaborada pelo autor (2026).

Nota: teste de McNemar exato pareado entre *zero-shot* e *few-shot*: $p = 1{,}00$ no conjunto de 900 perguntas e $p = 0{,}45$ no de 732 perguntas.

A associação entre a complexidade da consulta e a ocorrência de erro foi avaliada pelo teste exato de Fisher-Freeman-Halton. No conjunto de 900 perguntas, os erros por nível (simples, detalhada e complexa) passaram de 0, 3 e 6 na condição *zero-shot* para 1, 3 e 4 na condição *few-shot*, e a associação, significativa sem exemplos ($p = 0{,}0491$), deixou de sê-lo com a recuperação de exemplos ($p = 0{,}5490$). No conjunto de 732 perguntas, os erros passaram de 0, 2 e 4 para 0, 0 e 3,

e a associação permaneceu significativa nas duas condições (p = 0,0416 e p = 0,0258), mas por razão favorável: a RAG eliminou os erros dos níveis simples e detalhado, de modo que os três erros residuais se concentraram, por construção, no nível complexo. A escolha do teste foi decisiva nesse conjunto: pelo teste de qui-quadrado, cuja aproximação assintótica é inadequada com frequência esperada mínima de 1,78, a condição *zero-shot* resultaria não significativa (p = 0,0796). O V de Cramér variou de 0,044 a 0,099 nas quatro condições, o que indica efeito desprezível, e, com apenas 3 a 9 erros por condição, esses testes têm baixo poder, razão pela qual os valores de p não devem ser lidos como evidência de relação forte. Em síntese, a RAG não elevou a acurácia, mas alterou o perfil de erros: reduziu os erros nos níveis mais difíceis, introduziu um erro no nível simples do conjunto de 900 perguntas e, no conjunto de 732 perguntas, confinou os erros residuais ao nível complexo. A Figura 7 sintetiza, em painel comparativo, a acurácia, o custo, a duração, o consumo de *tokens* e a distribuição de erros das quatro condições.

Figura 7 – Síntese comparativa das quatro condições experimentais: (a) acurácia de execução; (b) custo variável por pergunta; (c) custo variável total; (d) duração total do experimento; (e) média de *tokens* por consulta; (f) erros por taxonomia



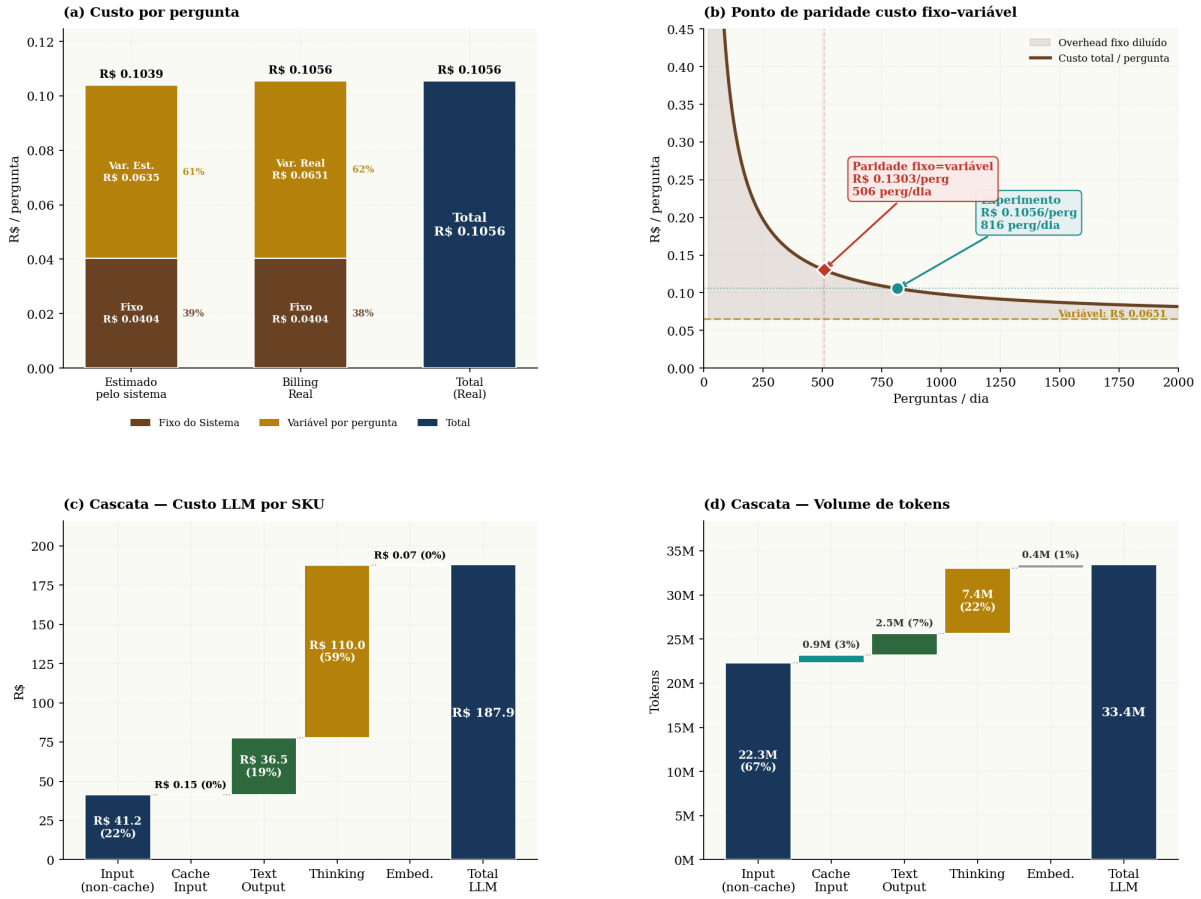
Fonte: Elaborada pelo autor (2026).

Cabe ressalva interpretativa quanto à comparação com a literatura. Os valores de acurácia obtidos superam amplamente os reportados em *benchmarks* abertos de *Text-to-SQL*, no entanto, refletem um regime de domínio especializado (*in-domain*): o sistema opera sobre um único esquema de banco de dados, com vocabulário operacional consolidado e camada de regras de negócio, condição substancialmente distinta da avaliação interdomínio sobre esquemas não vistos que caracteriza *benchmarks* como o *Spider* (Yu *et al.*, 2018), o *BIRD* (Li *et al.*, 2023) e o *Spider* 2.0 (Lei *et al.*, 2025). O desempenho deve, portanto, ser lido como evidência de eficácia da especialização da arquitetura para o domínio canavieiro, e não como acurácia diretamente comparável à desses *benchmarks* abertos. Convém ainda distinguir a acurácia de execução do sistema, que mede a proporção de perguntas convertidas em consulta SQL correta, do coeficiente de determinação (R^2) do modelo preditivo de consumo apresentado na seção 3.3, responsável por medir a precisão de uma regressão: por avaliarem objetos distintos, as duas medidas não são comparáveis entre si.

4.2 CUSTO OPERACIONAL E VALIDAÇÃO CONTRA A FATURA REAL

O custo variável total das 3.264 execuções foi de R$ 212,63, correspondente a 60% do custo total cobrado no período, o que resultou em um custo variável de R$ 0,0651 por pergunta (US$ 0,011). A decomposição do custo do modelo de linguagem por componente de cobrança evidenciou que os *tokens* de raciocínio (*thinking*) responderam pela maior parcela do custo variável, ao lado dos *tokens* de entrada, conforme a Figura 8.

Figura 8 – Estrutura econômica do sistema: (a) custo por pergunta estimado pelo sistema versus fatura real, com composição de custos fixos e variáveis; (b) ponto de paridade entre custo fixo e variável; (c) cascata de custo do modelo de linguagem por componente de cobrança; (d) cascata de volume de *tokens*



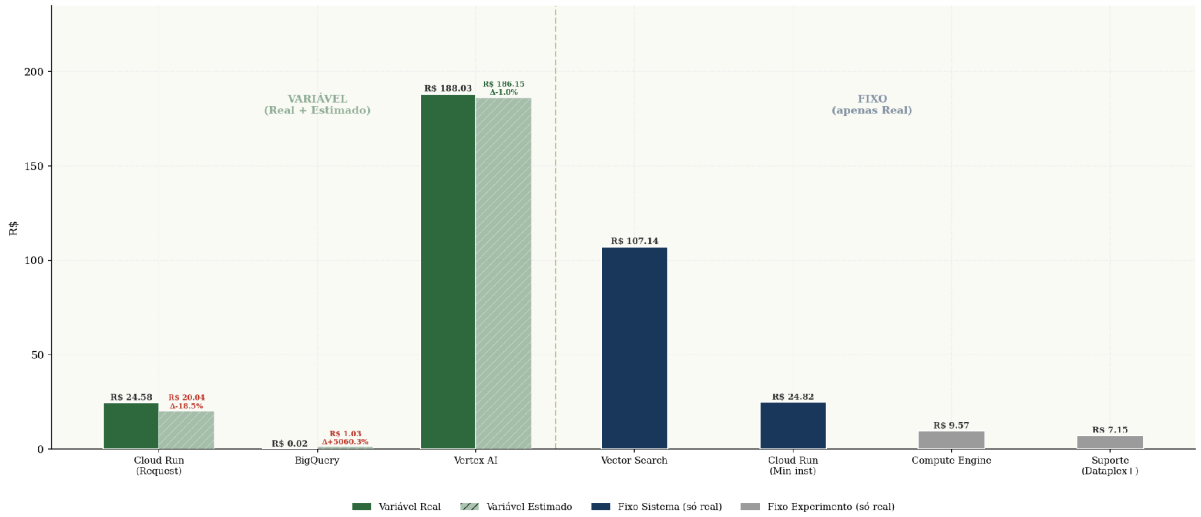
Fonte: Elaborada pelo autor (2026).

Uma das contribuições originais do trabalho é a validação cruzada do estimador de custo em tempo real contra a fatura real da nuvem. O custo agregado estimado pelo sistema apresentou desvio de apenas −2,5% frente ao valor efetivamente cobrado, e a análise por serviço (Figura 9) confirmou aderência elevada, com desvio de apenas −1,0% no serviço de inferência, o de maior peso na composição. A única discrepância relativa expressiva ocorreu no armazém de dados analítico (*BigQuery*), cujo custo estimado (R$ 1,03) superou o cobrado (R$ 0,02) em 5.060,3%. A divergência é de natureza metodológica, e não falha de medição: o estimador precifica todos os bytes processados pela tarifa cheia de consulta sob demanda, mas o provedor tarifa apenas o que excede a cota mensal gratuita de processamento. Como o experimento consumiu pouco além dessa cota, a fatura registrou valor residual. Em termos absolutos, a diferença corresponde a R$ 1,01, ou 0,3% do total cobrado no período, e não compromete a aderência agregada.

Reconhece-se, ainda assim, a limitação: uma versão futura do estimador pode contabilizar o consumo acumulado no mês e aplicar a tarifa somente à parcela que exceder a cota gratuita, o que tornaria a estimativa fiel também em regimes de baixo volume. Essa validação estabelece um instrumento auditável de gestão financeira por consulta, alinhado ao princípio de custo por resposta correta (Chen; Zaharia; Zou, 2023) e à rastreabilidade preconizada pela norma ISO/IEC 42001 (Thiers; Harned, 2024).

Desvio de sentido oposto ocorreu no serviço de execução conteinerizada, cujo custo estimado (R$ 20,04) ficou 18,5% abaixo do cobrado (R$ 24,58). A origem também é metodológica e decorre do ponto de instrumentação: o estimador deriva o custo de execução da duração do ciclo PDCA registrada na tabela de registros de interação, e esse registro é gravado exclusivamente pela requisição que processa a fila assíncrona. As demais requisições que cada pergunta gera no mesmo serviço, a recepção da mensagem pelo *webhook*, que apenas enfileira a tarefa, e o par de requisições do fluxo de validação humana, são faturadas pelo provedor, mas não produzem registro e, portanto, não entram na estimativa. Acresce que o serviço fatura tempo de instância alocada, e não tempo de requisição, de modo que os intervalos ociosos entre perguntas consecutivas não encontram correspondência no somatório das latências medidas. Em termos absolutos, a diferença é de R$ 4,54, ou 2,1% do custo variável do período. A correção é direta e fica indicada como refinamento futuro: estender a instrumentação às requisições de recepção e de validação, ou calibrar o estimador pelo tempo de instância alocada informado pelo próprio faturamento.

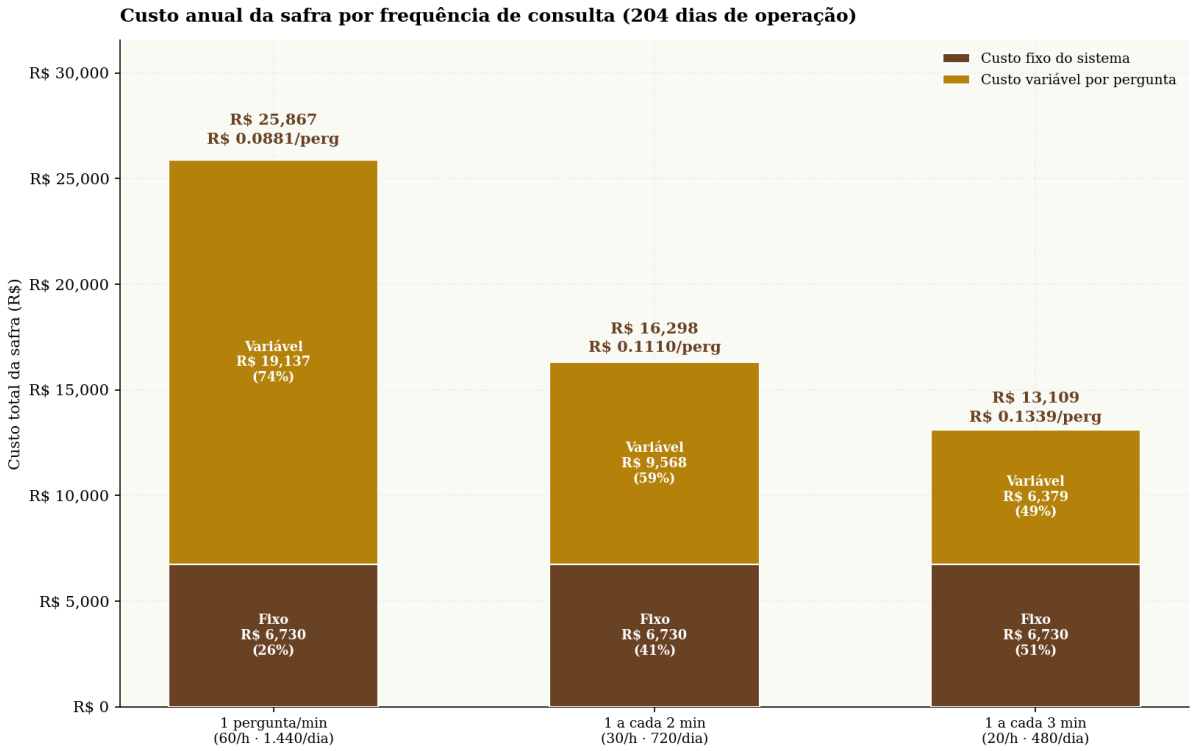
Figura 9 – Validação do estimador de custo contra a fatura real, por serviço de nuvem, com a separação das parcelas variáveis (dependentes do volume de consultas) e fixas



Fonte: Elaborada pelo autor (2026).

O custo fixo do sistema, dominado pelo *endpoint* de busca vetorial (R$ 107,14, equivalente a 81% do custo fixo), totalizou R$ 131,96 no período. A relação entre custo fixo e variável define o ponto de paridade em aproximadamente 507 consultas por dia, volume a partir do qual a parcela do custo variável passa a predominar sobre a parcela do custo fixo. Esse limiar tem leitura direta de escala: abaixo dele, o custo fixo de infraestrutura é diluído por poucas consultas, o que eleva o custo unitário; acima dele, situação compatível com o crescimento da frota e do número de supervisores ativos, o custo por pergunta cai expressivamente pela diluição dos custos estruturais. Esse custo fixo, contudo, mostra-se superdimensionado para o volume de exemplos utilizado: removendo o *endpoint* de busca vetorial (R$ 107,14), o custo fixo cairia de R$ 131,96 para cerca de R$ 24,82, o que deslocaria o ponto de paridade de 507 para aproximadamente 96 consultas por dia. A projeção para uma safra de 204 dias operacionais (Figura 10), em três cenários de carga, situou o custo total anual entre R$ 13.109 e R$ 25.867, com o custo por pergunta variando de R$ 0,134 a R$ 0,088 conforme a frequência de uso, o que evidencia que a economia de escala favorece operações de maior porte.

Figura 10 – Projeção do custo anual da safra (204 dias operacionais) por frequência de consulta, demonstrando a composição entre custos fixos e variáveis em três cenários de carga



Fonte: Elaborada pelo autor (2026).

Para dimensionar essa ordem de grandeza, o custo anual do sistema, de R$ 13.109 a R$ 25.867 por safra, equivale a uma faixa entre 0,12% e 0,25% do dispêndio operacional total, estimado em cerca de R$ 10,5 milhões (seção 3.1). Essa proporção reforça a favorabilidade da relação custo-benefício da arquitetura proposta como instrumento de apoio à decisão.

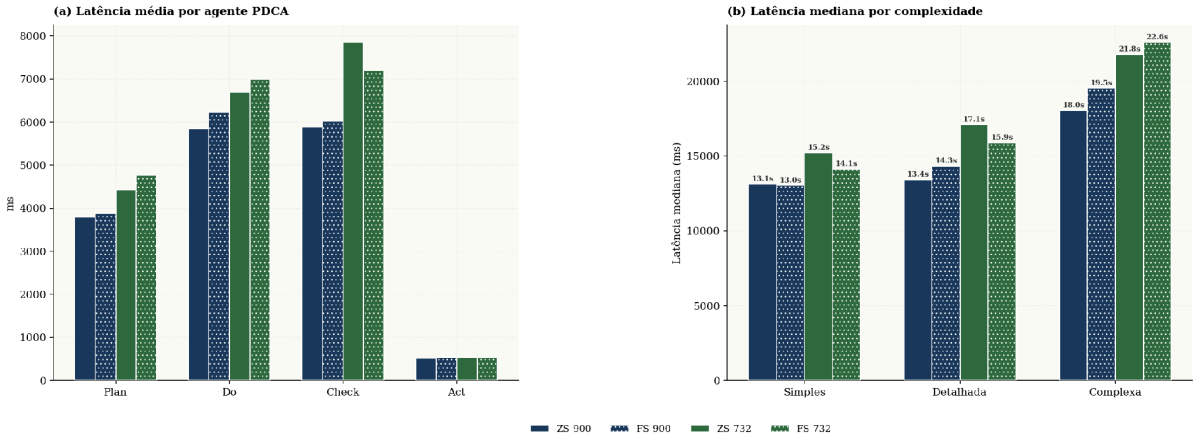
4.3 LATÊNCIA DO CICLO PDCA

A análise da latência por agente (Figura 11) evidenciou que os agentes Fazer (*Do*) e Checar (*Check*) concentram a maior parte do tempo de processamento: o agente Fazer (*Do*) variou de 5.373 ms a 5.839 ms entre as condições *zero-shot* e *few-shot*, e o agente Checar (*Check*), de 4.715 ms a 4.816 ms, ao passo que o agente Agir (*Act*) demandou cerca de 500 ms e o agente Planejar (*Plan*), aproximadamente 3.300 ms. A latência mediana total variou entre 14.088 ms (ZS-900) e 16.782 ms (ZS-732), com o teste de *Mann-Whitney U* identificando diferenças significativas entre as condições sem exemplos *zero-shot* e com

exemplos *few-shot* (p = 0,021 no conjunto de 900 perguntas; p < 0,001 no de 732 perguntas). O gargalo de latência concentra-se, portanto, nos agentes *Fazer* e *Checar*, o que sugere, como direção de otimização futura, uma redistribuição mais equilibrada de responsabilidades entre os agentes do ciclo PDCA. Um exemplo direto é transferir do agente Checar para o agente Agir a etapa de formatação da mensagem de resposta, hoje acumulada com a execução e a validação da consulta. O agente Agir consome cerca de 500 ms e limita-se à verificação de conformidade e ao envio, de modo que teria folga para absorver essa tarefa, enquanto o Checar ficaria restrito a executar a consulta e conferir a aderência dos dados à pergunta. Essa proposta é retomada na seção 4.7.

No eixo da complexidade da consulta (Figura 11b), a latência mediana cresceu de forma monotônica entre os três níveis definidos na seção 3.4.3, em todas as quatro condições experimentais. As consultas simples foram resolvidas entre 13,0 s e 15,2 s, as detalhadas entre 13,4 s e 17,1 s, e as complexas entre 18,0 s e 22,6 s, sendo este último o nível que concentrou os maiores tempos de resposta. O teste de Kruskal-Wallis confirmou que o efeito da complexidade sobre a latência foi estatisticamente significativo nas quatro condições (p < 0,001), com a estatística H variando de 143,70 na abordagem *zero-shot* do conjunto de 900 perguntas a 257,38 na modalidade *few-shot* do conjunto de 732 perguntas. Esse padrão é coerente com o gargalo identificado nos agentes Fazer e Checar (Figura 11a), no qual consultas mais complexas demandam um maior número de *tokens* de raciocínio na formulação e na validação do SQL, o que eleva o tempo de processamento. Observou-se, ainda, que a estratégia *few-shot* acentuou a diferenciação da latência entre os níveis, com H superior ao do *zero-shot* em ambos os conjuntos (194,40 contra 143,70 no grupo de 900, e 257,38 contra 143,98 no de 732), pois os exemplos recuperados pela RAG acrescentam *tokens* de entrada que se somam ao esforço já elevado das consultas complexas. Assim, embora a RAG tenha reduzido os erros nos níveis mais difíceis (seção 4.1), seu custo em latência recai sobretudo sobre as consultas mais elaboradas.

Figura 11 – Latência do ciclo PDCA: (a) tempo médio de processamento por agente nas quatro condições experimentais; (b) latência mediana por nível de complexidade da consulta (simples, detalhada e complexa)

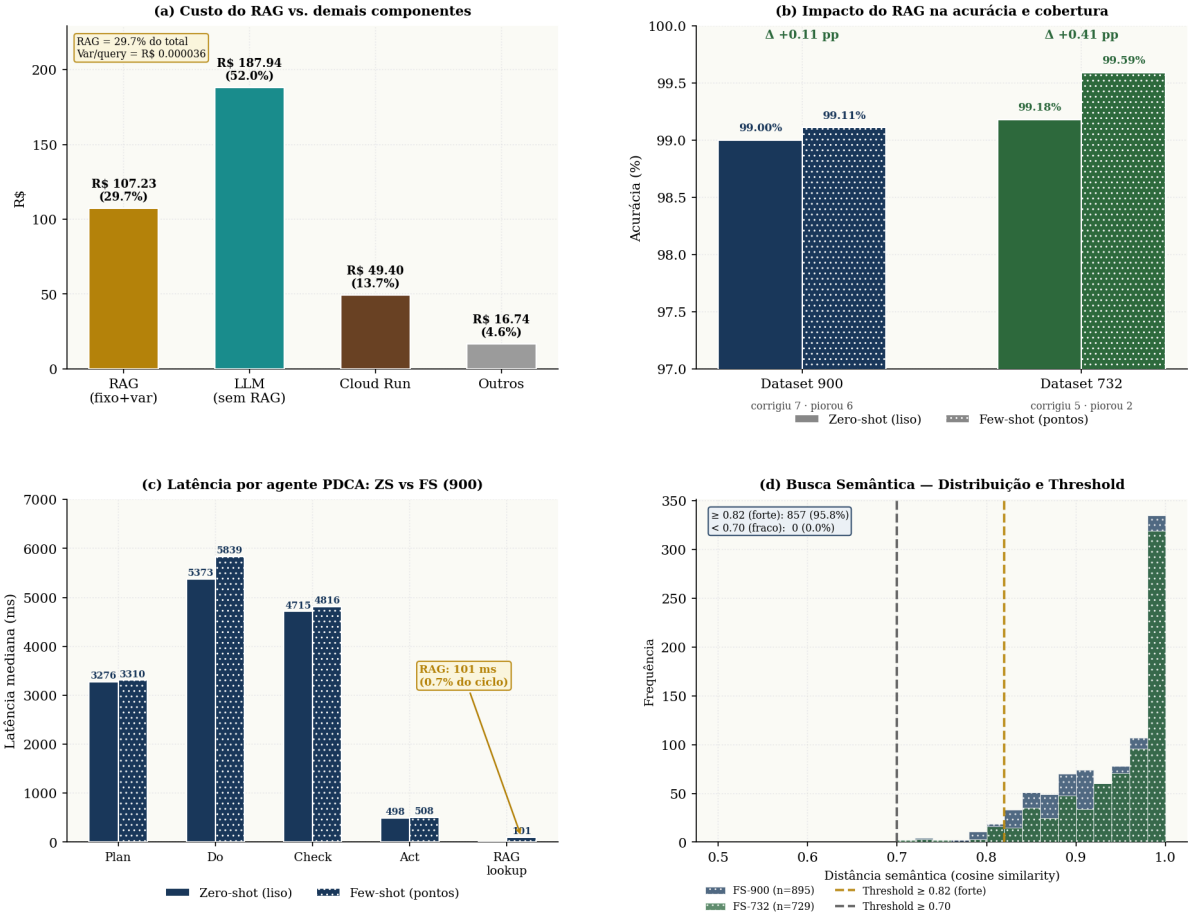


Fonte: Elaborada pelo autor (2026).

4.4 DESEMPENHO DA GERAÇÃO AUMENTADA POR RECUPERAÇÃO

A Figura 12 apresenta a análise integrada do componente de Geração Aumentada por Recuperação. O componente concentrou R$ 107,23 do custo total, o que equivale a 29,7% do orçamento global do período, com impacto majoritariamente fixo, em razão principalmente do *endpoint* de busca vetorial do *Vertex AI*. Quanto ao impacto na acurácia, o ganho líquido foi de +0,11 ponto percentual no conjunto de 900 perguntas (sete consultas corrigidas e seis degradadas) e de +0,41 ponto percentual no de 732 perguntas (cinco corrigidas e duas degradadas). A cobertura da busca semântica foi próxima de 100%: 95,8% das consultas recuperaram exemplos de forte proximidade (similaridade por cosseno ≥ 0,82) e nenhuma recuperou exemplos de baixa similaridade (< 0,70), com a latência de recuperação de apenas 101 ms, o que correspondeu a 0,7% do ciclo total.

Figura 12 – Desempenho do componente de Geração Aumentada por Recuperação: (a) custo da RAG frente aos demais componentes; (b) impacto na acurácia e cobertura por conjunto; (c) latência por agente, com a recuperação representando 0,7% do ciclo; (d) distribuição da similaridade semântica e limiares de recuperação



Fonte: Elaborada pelo autor (2026).

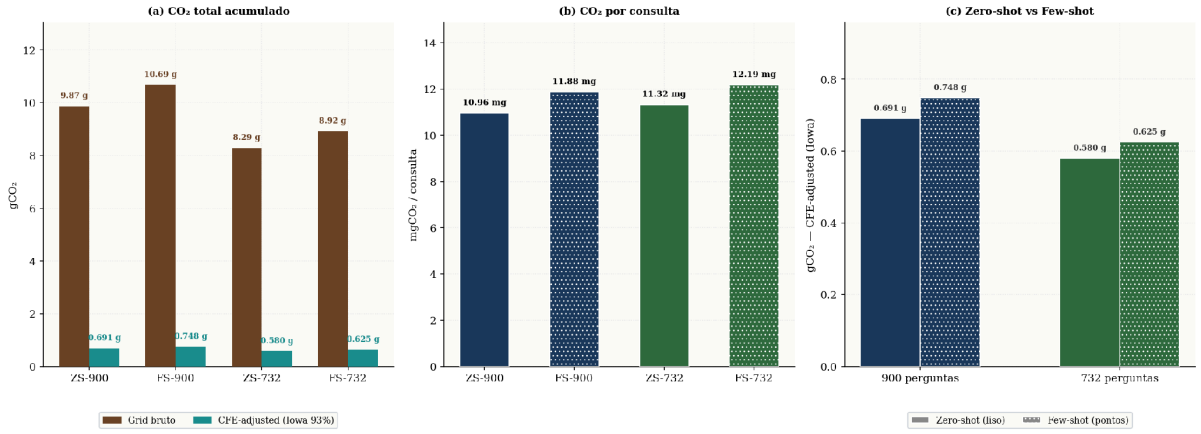
Como o índice de busca vetorial opera por força bruta sobre apenas 90 vetores, seu custo decorre da disponibilidade do *endpoint* dedicado, e não do processamento da busca em si. Por se tratar de um componente predominantemente fixo, cujo maior valor está na alteração do perfil de erros entre os níveis de complexidade da consulta (seção 4.1), e não no ganho incremental de acurácia, delineiam-se duas direções para pesquisas futuras. A primeira consiste na flexibilização da engenharia de *prompt,* com redução do número de *tokens* de entrada e de raciocínio, obtida por dois mecanismos complementares. O primeiro mecanismo baseia-se em ampliar a base de exemplos validados, de modo que a recuperação de casos semelhantes assuma parte do esforço hoje concentrado na construção manual das instruções recebidas por cada agente. O segundo envolve migrar a base de regras de negócio, atualmente injetada de forma estática e integral

em cada consulta, para recuperação semântica sob demanda, de maneira que apenas as regras pertinentes a cada pergunta componham o contexto, em vez de sobrecarregar o *prompt* com o conjunto completo a cada execução. A segunda direção é a migração do *endpoint* de busca vetorial para o modo sem servidor *(serverless)*, o que eliminaria a maior parcela do custo fixo do sistema.

4.5 PEGADA DE CARBONO DA INFERÊNCIA

A pegada de carbono foi estimada pela Equação 1 (Figura 13). No cenário de matriz elétrica bruta, as emissões totais por condição variaram entre 8,29 g (ZS-732) e 10,69 g (FS-900); no cenário ajustado por energia livre de carbono, as emissões totais caíram para a faixa de 0,58 g a 0,75 g por condição, o que corresponde a uma emissão média entre 0,77 mg e 0,85 mg de CO_2 por consulta. Para contextualizar, seriam necessárias cerca de 1,30 milhão de consultas para gerar 1 kg de CO_2 no cenário ajustado, magnitude desprezível frente ao potencial de redução de emissões no transporte canavieiro, no qual ganhos marginais de eficiência da frota se traduzem em reduções expressivas de combustível e de gases de efeito estufa (Bartholomeu; Péra; Caixeta-Filho, 2016).

Figura 13 – Pegada de carbono da inferência nas quatro condições experimentais: (a) emissões de CO_2 totais acumuladas nos cenários bruto e ajustado por energia livre de carbono; (b) emissão média de CO_2 por consulta; (c) comparação entre as abordagens *zero-shot* e *few-shot*



Fonte: Elaborada pelo autor (2026).

4.6 CATEGORIZAÇÃO DE ERROS E CONFORMIDADE COM A LGPD

A distribuição de erros segundo a taxonomia tripartida revelou um achado de elevado significado prático: zero erros de vinculação ao esquema da tabela do BigQuery (*schema linking*) em 3.264 execuções (Tabela 3). Os 26 erros observados distribuíram-se igualmente entre lógica SQL (13 ocorrências) e interpretação semântica (13 ocorrências). A análise qualitativa dos erros de lógica SQL apontou como padrões predominantes a inversão de ordenação e o recorte temporal incorreto, ao passo que as falhas de interpretação decorreram, sobretudo, de classificações incompletas da intenção na fase de planejamento. A ausência total de erros de *schema linking* evidencia a eficácia combinada das camadas de regras de negócio, como também da recuperação de exemplos validados na ancoragem dos termos da pergunta ao esquema real.

Tabela 3 – Distribuição de erros por tipo, segundo a taxonomia de Yu *et al.* (2018) e Lei *et al.* (2025), e por condição experimental

Condição	(i) *Schema linking*	(ii) Lógica SQL	(iii) Interpretação	Total
ZS-900	0	6	3	9
FS-900	0	5	3	8
ZS-732	0	2	4	6
FS-732	0	0	3	3
Total	**0**	**13**	**13**	**26**

Fonte: Elaborada pelo autor (2026).

Quanto à conformidade regulatória, durante as 3.264 execuções, foram registrados zero bloqueios pela camada de verificação de dados pessoais identificáveis implementada no agente Agir (*Act*), o que confirmou a eficácia da prevenção por concepção (*privacy by design*), distribuída entre os agentes Planejar, Fazer e Checar. Do ponto de vista agronômico, a disponibilização instantânea de indicadores de consumo, eficiência e custo de transporte via mensageria tem implicação direta na velocidade da ação corretiva: ao traduzir o vocabulário operacional do supervisor diretamente em consultas, sem intermediação de analistas, o sistema reduz o intervalo entre a identificação de um desvio e a intervenção em campo.

4.7 SÍNTESE, LIMITAÇÕES E DIREÇÕES FUTURAS

O valor agronômico da arquitetura reside menos na geração da consulta em si e mais na velocidade com que ela entrega ao tomador de decisão a informação necessária para perseguir ganhos marginais de eficiência. A latência de 14 s a 17 s por consulta deve ser lida em contraste com o fluxo manual que o sistema substitui: abrir um painel, aplicar filtros e interpretar uma planilha consomem, na prática, alguns minutos. São decisões operacionais cotidianas, como qual caminhão enviar a cada zona e como reduzir o consumo por rota, que o sistema passa a subsidiar em tempo real, a um custo computacional muito inferior ao dispêndio de combustível potencialmente evitável.

Três contribuições replicáveis emergem do trabalho. A primeira é metodológica: o protocolo de validação que integra o ciclo PDCA, o delineamento fatorial 2×2 e os testes exatos de McNemar e de Fisher-Freeman-Halton constitui processo replicável para sistemas *Text-to-SQL* aplicados a dados agrícolas, alinhado às diretrizes da ISO/IEC 42001. A segunda é econômica: a validação cruzada do estimador de custo contra a fatura do provedor estabelece instrumento auditável de gestão de custo por consulta, o que evidencia potencial de economia e eficiência orçamentária, e não a viabilidade financeira do empreendimento como um todo, cuja avaliação demandaria análise de investimento específica. A terceira é regulatória: a conformidade nativa com a LGPD, com zero bloqueios em 3.264 execuções, confirma a proteção por concepção como modelo replicável para o agronegócio.

Reconhecem-se limitações que circunscrevem a generalização dos resultados. A avaliação foi conduzida sobre um único esquema de banco de dados e uma única frota, o que caracteriza regime de domínio especializado e impede a transposição direta a outras operações sem validação específica. O estimador de custo, por sua vez, não modela a cota gratuita de processamento analítico nem as requisições de recepção e de validação, limitações quantificadas na seção 4.2.

As direções de pesquisa futura organizam-se em três frentes. No custo, a migração do *endpoint* de busca vetorial para o modo sem servidor reduziria 81% do custo fixo e deslocaria o ponto de paridade de 507 para cerca de 96 consultas por dia. Na arquitetura, a transferência da formatação da resposta do agente Checar para o agente Agir distribuiria a carga e atacaria o gargalo de latência, o mesmo valendo para a migração das regras de negócio de injeção estática para

recuperação semântica sob demanda, com redução do número de *tokens* de entrada. No modelo, a comparação direta com modelos de linguagem de menor porte (*Small Language Models*, SLM), cujo custo de inferência tende a ser uma ordem de grandeza inferior, permitiria avaliar o compromisso entre custo, acurácia e latência em configurações mais enxutas e reduzir o ponto de paridade entre custo fixo e variável, condição para a adoção em operações de menor porte, historicamente limitadas pela baixa maturidade analítica (Tyrychtr; Ulman; Vostrovský, 2015).

5 CONCLUSÃO

O sistema multiagente de inteligência artificial generativa, organizado segundo o ciclo PDCA e disponibilizado por mensageria, converte perguntas de linguagem natural em consultas SQL corretas sobre dados de telemetria do transporte canavieiro, com acurácia de execução entre 99,00% e 99,59%, custo variável de R$ 0,0651 por pergunta e conformidade com a LGPD verificada em todas as 3.264 execuções.

A hipótese central não foi confirmada. A Geração Aumentada por Recuperação não elevou significativamente a acurácia, porque a geração sem exemplos já operava próxima ao limite superior de desempenho. Seu efeito manifestou-se de outra forma, como alteração do perfil de erros: os erros diminuíram nos níveis mais difíceis da consulta SQL, surgiu um erro no nível simples e, no conjunto de 732 perguntas, os erros residuais ficaram restritos ao nível complexo. Como o tamanho de efeito da associação entre complexidade e erro foi desprezível em todas as condições (V de Cramér $\leq 0{,}099$), o resultado deve ser lido como mudança na localização dos erros, e não como mecanismo de estabilização demonstrado.

Os quatro objetivos específicos foram atendidos. A arquitetura sem servidor foi projetada, implementada e sustentou o experimento completo a partir de dados reais de 30.791 viagens; o efeito da RAG foi quantificado sob delineamento fatorial 2×2; o estimador de custo em tempo real foi validado contra o faturamento do provedor, com desvio agregado de −2,5%; e a pegada de carbono da inferência e a conformidade regulatória foram aferidas em todas as execuções.

A adoção do sistema por outras unidades sucroenergéticas mostra-se viável, condicionada à validação em cada contexto operacional.

REFERÊNCIAS

AGRESTI, A. **Categorical data analysis**. 2nd ed. Hoboken: John Wiley & Sons, 2002. E-book. DOI: 10.1002/0471249688. Disponível em: https://doi.org/10.1002/0471249688. Acesso em: 18 set. 2026.

BARTHOLOMEU, D. B.; PÉRA, T. G.; CAIXETA-FILHO, J. V. Logística sustentável: avaliação de estratégias de redução das emissões de CO_2 no transporte rodoviário de cargas. **Journal of Transport Literature**, São José dos Campos, v. 10, n. 3, p. 15-19, jul. 2016. DOI: 10.1590/2238-1031.jtl.v10n3a3. Disponível em: https://www.scielo.br/j/jtl/a/hmCRKy5mgYswyFvrKRCTzvG/. Acesso em: 18 set. 2026.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2026]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 18 set. 2026.

BROWN, T. *et al.* Language models are few-shot learners. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, [S. l.]. **Advances in neural information processing systems**. Red Hook: Curran Associates, 2020. v. 33, p. 1877-1901. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html. Acesso em: 21 set. 2026.

CHEN, L.; ZAHARIA, M.; ZOU, J. **FrugalGPT**: how to use large language models while reducing cost and improving performance. [S. l.]: arXiv, 2023. Preprint arXiv:2305.05176. DOI: 10.48550/arXiv.2305.05176. Disponível em: https://doi.org/10.48550/arXiv.2305.05176. Acesso em: 18 set. 2026.

COHEN, J. **Statistical power analysis for the behavioral sciences**. 2nd ed. Hillsdale: Lawrence Erlbaum Associates, 1988.

COMITÊ GESTOR DA INTERNET NO BRASIL. **Pesquisa sobre o uso das tecnologias de informação e comunicação nos domicílios brasileiros**: TIC Domicílios 2025. São Paulo: CGI.br, 2026. E-book. ISBN 978-65-6165-016-8. Disponível em: https://cetic.br/pt/publicacao/pesquisa-sobre-o-uso-das-tecnologias-de-informacao-e-comunicacao-nos-domicilios-brasileiros-tic-domicilios-2025/. Acesso em: 13 set. 2026.

FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **FAOSTAT: crops and livestock products**. Rome: FAO, 2024. Base de dados. Disponível em: https://www.fao.org/faostat/en/#data/QCL. Acesso em: 18 set. 2026.

FREEMAN, G. H.; HALTON, J. H. Note on an exact treatment of contingency, goodness of fit and other problems of significance. **Biometrika**, Oxford, v. 38, n. 1/2, p. 141-149, June 1951. DOI: 10.1093/biomet/38.1-2.141. Disponível em: https://doi.org/10.1093/biomet/38.1-2.141. Acesso em: 3 out. 2026.

GAMAGE, G. *et al.* Multi-agent RAG chatbot architecture for decision support in net-zero emission energy systems. In: IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL TECHNOLOGY, 25., 2024, Bristol. **Proceedings** [...]. Piscataway: IEEE, 2024. p. 1-6. DOI: 10.1109/ICIT58233.2024.10540920. Disponível em: https://doi.org/10.1109/ICIT58233.2024.10540920. Acesso em: 18 set. 2026.

GAO, D. *et al.* Text-to-SQL empowered by large language models: a benchmark evaluation. **Proceedings of the VLDB Endowment**, New York, v. 17, n. 5, p. 1132-1145, 2024. DOI: 10.14778/3641204.3641221. Disponível em: https://doi.org/10.14778/3641204.3641221. Acesso em: 18 set. 2026.

GAO, Y. *et al.* **Retrieval-augmented generation for large language models**: a survey. [S. l.]: arXiv, 2024. Preprint arXiv:2312.10997. DOI: 10.48550/arXiv.2312.10997. Disponível em: https://doi.org/10.48550/arXiv.2312.10997. Acesso em: 18 set. 2026.

GOOGLE CLOUD. **BigQuery explainable AI overview**. Mountain View: Google, 2026a. Disponível em: https://cloud.google.com/bigquery/docs/xai-overview. Acesso em: 22 jun. 2026.

GOOGLE CLOUD. **Exchange rates**. Mountain View: Google, 2026b. Disponível em: https://cloud.google.com/skus/exchange-rates. Acesso em: 22 jun. 2026.

GOOGLE CLOUD. **Introduction to ML in BigQuery**. Mountain View: Google, 2026c. Disponível em: https://cloud.google.com/bigquery/docs/bqml-introduction. Acesso em: 10 jun. 2026.

GUO, T. *et al.* Large language model based multi-agents: a survey of progress and challenges. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, 33., 2024, Jeju. **Proceedings** [...]. [S. l.]: IJCAI, 2024. p. 8048-8057. DOI: 10.24963/ijcai.2024/890. Disponível em: https://doi.org/10.24963/ijcai.2024/890. Acesso em: 18 set. 2026.

HOLLANDER, M.; WOLFE, D. A.; CHICKEN, E. **Nonparametric statistical methods**. 3rd ed. Hoboken: John Wiley & Sons, 2014. E-book. DOI: 10.1002/9781119196037. Disponível em: https://doi.org/10.1002/9781119196037. Acesso em: 18 set. 2026.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE. **2006 IPCC guidelines for national greenhouse gas inventories**. Hayama: IGES, 2006. Disponível em: https://www.ipcc-nggip.iges.or.jp/public/2006gl/. Acesso em: 18 set. 2026.

KAMILARIS, A.; KARTAKOULLIS, A.; PRENAFETA-BOLDÚ, F. X. A review on the practice of big data analysis in agriculture. **Computers and Electronics in Agriculture**, Amsterdam, v. 143, p. 23-37, Dec. 2017. DOI: 10.1016/j.compag.2017.09.037. Disponível em: https://doi.org/10.1016/j.compag.2017.09.037. Acesso em: 18 set. 2026.

KIANGALA, K. S.; WANG, Z. An experimental hybrid customized AI and generative AI chatbot human machine interface to improve a factory troubleshooting downtime

in the context of Industry 5.0. **The International Journal of Advanced Manufacturing Technology**, London, v. 132, n. 5/6, p. 2715-2733, 2024. DOI: 10.1007/s00170-024-13492-0. Disponível em: https://doi.org/10.1007/s00170-024-13492-0. Acesso em: 18 set. 2026.

LEI, F. *et al.* Spider 2.0: evaluating language models on real-world enterprise text-to-SQL workflows. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 13., 2025, Singapore. **Proceedings** [...]. [S. l.]: ICLR, 2025. Disponível em: https://proceedings.iclr.cc/paper_files/paper/2025/hash/46c10f6c8ea5aa6f267bcdabcb123f97-Abstract-Conference.html. Acesso em: 18 set. 2026.

LEWIS, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, [S. l.]. **Advances in neural information processing systems**. Red Hook: Curran Associates, 2020. v. 33, p. 9459-9474. Disponível em: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Acesso em: 21 set. 2026.

LI, J. *et al.* Can LLM already serve as a database interface? A big bench for large-scale database grounded text-to-SQLs. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans. **Advances in neural information processing systems**. Red Hook: Curran Associates, 2023. v. 36, p. 42330-42357. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/hash/83fc8fab1710363050bbd1d4b8cc0021-Abstract-Datasets_and_Benchmarks.html. Acesso em: 21 set. 2026.

LUNDBERG, S. M.; LEE, S. I. A unified approach to interpreting model predictions. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach. **Advances in neural information processing systems**. Red Hook: Curran Associates, 2017. v. 30, p. 4765-4774. Disponível em: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html. Acesso em: 21 set. 2026.

MOEN, R. D.; NORMAN, C. L. Evolution of the PDCA cycle. In: ANQ CONGRESS, 7., 2009, Tokyo. **Proceedings** [...]. Tokyo: Asian Network for Quality, 2009.

PATTERSON, D. *et al.* The carbon footprint of machine learning training will plateau, then shrink. **Computer**, Los Alamitos, v. 55, n. 7, p. 18-28, July 2022. DOI: 10.1109/MC.2022.3148714. Disponível em: https://doi.org/10.1109/MC.2022.3148714. Acesso em: 18 set. 2026.

PECEGE. Custos de produção de cana-de-açúcar, açúcar e etanol na safra 25/26. Piracicaba: Pecege Consultoria e Projetos, 2026. Trabalho apresentado na Expedição Custos Cana 2026. Disponível em: https://www.custoscana.com.br/. Acesso em: 21 set. 2026.

POURREZA, M.; RAFIEI, D. DIN-SQL: decomposed in-context learning of text-to-SQL with self-correction. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans. **Advances in neural**

information processing systems. Red Hook: Curran Associates, 2023. v. 36, p. 36339-36348. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/hash/72223cc66f63ca1aa59edaec1b3670e6-Abstract-Conference.html. Acesso em: 21 set. 2026.

RAY, P. P. Generative AI and its impact on sugarcane industry: an insight into modern agricultural practices. **Sugar Tech**, New Delhi, v. 26, n. 2, p. 325-332, 2024. DOI: 10.1007/s12355-023-01358-w. Disponível em: https://doi.org/10.1007/s12355-023-01358-w. Acesso em: 18 set. 2026.

SEABOLD, S.; PERKTOLD, J. Statsmodels: econometric and statistical modeling with Python. In: PYTHON IN SCIENCE CONFERENCE, 9., 2010, Austin. **Proceedings** [...]. Austin: SciPy, 2010. p. 92-96. DOI: 10.25080/Majora-92bf1922-011. Disponível em: https://doi.org/10.25080/Majora-92bf1922-011. Acesso em: 18 set. 2026.

SHAPLEY, L. S. A value for n-person games. In: KUHN, H. W.; TUCKER, A. W. (ed.). **Contributions to the theory of games**. Princeton: Princeton University Press, 1953. v. 2, p. 307-317. DOI: 10.1515/9781400881970-018. Disponível em: https://doi.org/10.1515/9781400881970-018. Acesso em: 18 set. 2026.

SILVA, R. P. da *et al.* Controle estatístico aplicado ao processo de colheita mecanizada de cana-de-açúcar. **Engenharia Agrícola**, Jaboticabal, v. 28, n. 2, p. 292-304, abr./jun. 2008. DOI: 10.1590/S0100-69162008000200010. Disponível em: https://www.scielo.br/j/eagri/a/8wFFWF378FcwksgfxBvLrhF/. Acesso em: 18 set. 2026.

THIERS, F. A.; HARNED, Z. **The emerging role of ISO 42001 certification in fostering the deployment of responsible generative AI healthcare solutions**. [S. l.]: OSF Preprints, 2024. Preprint. DOI: 10.31219/osf.io/5ks37. Disponível em: https://doi.org/10.31219/osf.io/5ks37. Acesso em: 18 set. 2026.

TYRYCHTR, J.; ULMAN, M.; VOSTROVSKÝ, V. Evaluation of the state of the business intelligence among small Czech farms. **Agricultural Economics – Czech**, Prague, v. 61, n. 2, p. 63-71, 2015. DOI: 10.17221/108/2014-AGRICECON. Disponível em: https://doi.org/10.17221/108/2014-AGRICECON. Acesso em: 18 set. 2026.

VASWANI, A. *et al.* Attention is all you need. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach. **Advances in neural information processing systems**. Red Hook: Curran Associates, 2017. v. 30, p. 5998-6008. Disponível em: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Acesso em: 21 set. 2026.

VIRTANEN, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature Methods**, London, v. 17, n. 3, p. 261-272, Mar. 2020. DOI: 10.1038/s41592-019-0686-2. Disponível em: https://doi.org/10.1038/s41592-019-0686-2. Acesso em: 18 set. 2026.

VOLTARELLI, M. A. *et al.* Failure mode and effect analysis (FMEA) in mechanized harvest of sugarcane billets. **Engenharia Agrícola**, Jaboticabal, v. 38, n. 1, p. 88-96, Jan./Feb. 2018. DOI: 10.1590/1809-4430-eng.agric.v38n1p88-96/2018. Disponível em: https://www.scielo.br/j/eagri/a/zM4pjfrQNsDYYbkSShh3SRk/. Acesso em: 18 set. 2026.

WOLFERT, S. *et al.* Big data in smart farming: a review. **Agricultural Systems**, Amsterdam, v. 153, p. 69-80, May 2017. DOI: 10.1016/j.agsy.2017.01.023. Disponível em: https://doi.org/10.1016/j.agsy.2017.01.023. Acesso em: 18 set. 2026.

YAN, L. *et al.* VizChat: enhancing learning analytics dashboards with contextualised explanations using multimodal generative AI chatbots. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE IN EDUCATION, 25., 2024, Recife. **Artificial intelligence in education**: proceedings, part II. Cham: Springer, 2024. p. 180-193. (Lecture Notes in Computer Science, 14830). DOI: 10.1007/978-3-031-64299-9_13. Disponível em: https://doi.org/10.1007/978-3-031-64299-9_13. Acesso em: 18 set. 2026.

YANG, B. *et al.* **AgriGPT**: a large language model ecosystem for agriculture. [S. l.]: arXiv, 2025. Preprint arXiv:2508.08632. DOI: 10.48550/arXiv.2508.08632. Disponível em: https://doi.org/10.48550/arXiv.2508.08632. Acesso em: 18 set. 2026.

YU, T. *et al.* Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2018, Brussels. **Proceedings** [...]. Stroudsburg: Association for Computational Linguistics, 2018. p. 3911-3921. DOI: 10.18653/v1/D18-1425. Disponível em: https://doi.org/10.18653/v1/D18-1425. Acesso em: 18 set. 2026.

APÊNDICE A – EXEMPLOS DE CONSULTAS POR NÍVEL DE COMPLEXIDADE

Este apêndice reúne exemplos representativos de perguntas em linguagem natural submetidas ao agente PDCA e das respectivas respostas entregues na interface do *Telegram*, organizados segundo os três níveis de complexidade da consulta SQL definidos na seção 3.4.3: simples (Figura 1A), detalhado (Figura 2A) e complexo (Figura 3A).

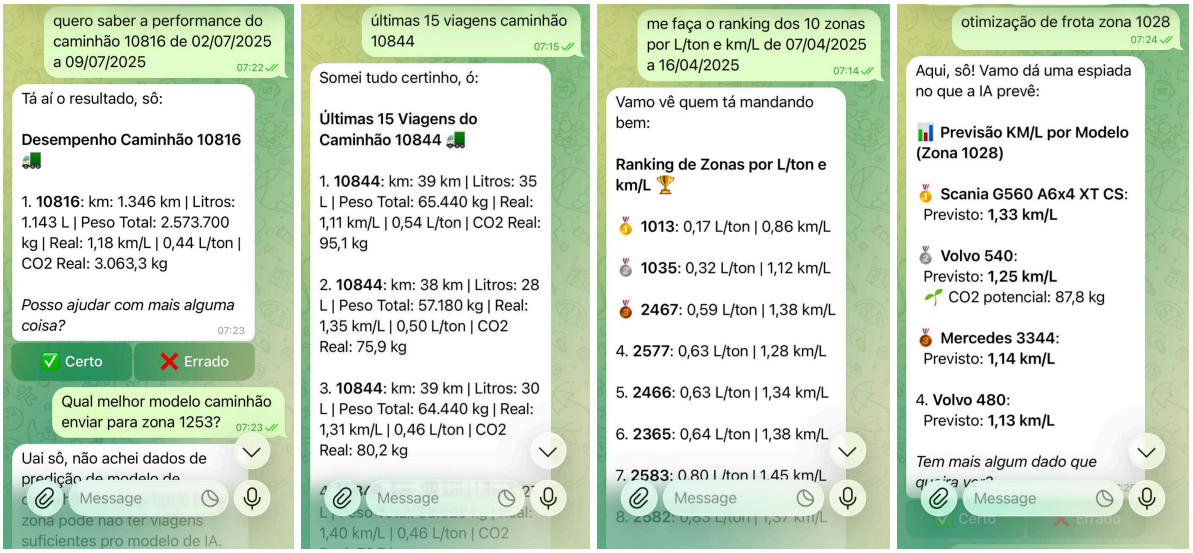
Figura 1A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível simples

Fonte: Elaborada pelo autor (2026).

Figura 2A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível detalhado

Fonte: Elaborada pelo autor (2026).

Figura 3A – Exemplos de perguntas em linguagem natural e respectivas respostas do agente PDCA, na interface do *Telegram*, para consultas de nível complexo



Fonte: Elaborada pelo autor (2026).

DADOS CURRICULARES

IDENTIFICAÇÃO	
	Lucas Eduardo Zonfrilli Nascido em Assis, SP, em 18 de outubro de 1992
Nacionalidade	brasileira
Nome em citações bibliográficas:	ZONFRILLI, Lucas Eduardo ZONFRILLI, L. E.
Currículo Lattes	http://lattes.cnpq.br/7260904166315536
ORCID	https://orcid.org/0009-0006-9673-371X
FORMAÇÃO ACADÊMICA	
2024/2026	Mestrado em Agronomia (Produção Vegetal) Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Câmpus de Jaboticabal
2021/2024	Tecnologia em Gestão da Tecnologia da Informação Faculdade de Tecnologia do Estado de São Paulo (FATEC), Assis
2011/2017	Bacharelado em Ciências Econômicas Universidade Estadual de Londrina (UEL)
2009/2010	Técnico em Administração Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS)
2024	Extensão universitária: Estágio de Férias no Laboratório de Máquinas e Mecanização Agrícola (180 h) Universidade Estadual Paulista (UNESP)
2023	Intercâmbio acadêmico em Engenharia de Redes Universidad Autónoma del Estado de Quintana Roo (UQROO), México
2023	Manutenção de Drones (2 h); Procedimentos Legais para Habilitação e Uso de Drones (2 h) Serviço Nacional de Aprendizagem Rural (SENAR-AR/SP)
2022/2023	Data Science no Agronegócio (420 h) Agroadvance
PRODUÇÃO BIBLIOGRÁFICA	

ZONFRILLI, L. E.; CAMOLESI, A. R.; CANATA, T. F. Impact of telemetry data preprocessing on the accuracy of fuel consumption predictive models in sugarcane transport. In: INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE, 17.; BRAZILIAN CONGRESS ON PRECISION AND DIGITAL AGRICULTURE, 11., 2026, Porto Alegre. Proceedings [...]. Monticello: International Society of Precision Agriculture, 2026.

ZONFRILLI, L. E. Investigación aplicada utilizando datos públicos del Ministerio de Salud de Brasil: análisis histórico del período de 2014 a 2022. In: JORNADAS DE HUMANIDADES EN LA FRONTERA MÉXICO-BELICE, 4., 2023, Chetumal. Chetumal: Universidad Autónoma del Estado de Quintana Roo, 2023. Evento virtual. (Apresentação de trabalho).

ZONFRILLI, L. E.; SOARES, M. S.; PEPECE, A. R.; CARDOSO, F. É. Aplicação Python para coletar dados públicos e aferir modelos estatísticos: pesquisa aplicada com dados públicos do Ministério da Saúde brasileiro. In: JORNADA CIENTÍFICA E TECNOLÓGICA, 5., 2023, Bauru. Bauru: Faculdade de Tecnologia de Bauru, 2023. (Apresentação de trabalho, artigo completo).

ZONFRILLI, L. E. Setor sucroenergético: panorama competitivo brasileiro e competitividade internacional, de 2008 a 2013. Orientador: Carlos Eduardo Caldarelli. 2016. Trabalho de Conclusão de Curso (Bacharelado em Ciências Econômicas) - Universidade Estadual de Londrina, Londrina, 2016.

COLABORAÇÃO EM TRABALHOS ORIENTADOS

Colaboração técnica em manutenção e funcionamento de fluxômetro, análise de dados e estatística, e revisão da redação.

Trabalho de conclusão de curso de graduação

LOPES, R. C. Consumo de combustível e emissão de CO_2 em dois sistemas de preparo do solo utilizando trator de pequeno porte. 2025. Trabalho de Conclusão de Curso (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

SILVA, R. H. S. Emissão de CO_2 do solo em três sistemas de preparo: avaliação comparativa e implicações para a sustentabilidade agrícola. 2025. Trabalho de Conclusão de Curso (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

Iniciação científica

ZIMERMAM, N. A. Interpolação espacial da produtividade e maturação do café para estimativa de retorno econômico. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

SANTOS, F. Z. Inteligência artificial generativa para consulta automatizada de manuais técnicos de máquinas agrícolas. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

DONINI, G. Avaliação do consumo de combustível de tratores de classe IV, em diferentes profundidades de sulcação da cana-de-açúcar, utilizando microcontroladores Arduino. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

MORCELLE, D. S. Sistema de monitoramento para consumo de combustível em máquinas agrícolas com microcontrolador Arduino. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

MONACHESI, F. P. Identificação de greening em citros a partir do processamento de imagens multiespectrais. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

SÁ, A. M. B. Sistema embarcado de monitoramento da distância, velocidade e patinagem de máquinas agrícolas com Arduino. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

SILVA, S. A. H. Caracterização da dependência espacial de atributos qualitativos da cana-de-açúcar. 2025. Iniciação Científica (Graduação em Engenharia Agronômica) - Universidade Estadual Paulista (UNESP), Jaboticabal, 2025.

PARTICIPAÇÃO EM EVENTOS CIENTÍFICOS

INTERNATIONAL CONFERENCE ON PRECISION AGRICULTURE, 17.; BRAZILIAN CONGRESS ON PRECISION AND DIGITAL AGRICULTURE, 11., 2026, Porto Alegre. Impact of telemetry data preprocessing on the accuracy of fuel consumption predictive models in sugarcane transport. 2026. (Congresso, apresentação oral).