RODRIGO EDIEL VOGT

VALIDAÇÃO AUTOMÁTICA DE RESPOSTAS TÉCNICAS AUTOMOTIVAS UTILIZANDO MODELOS DE LINGUAGEM E RECUPERAÇÃO DE INFORMAÇÃO

Trabalho de Conclusão de Curso apresentado ao curso de Ciência da Computação da Universidade Federal da Fronteira Sul (UFFS), como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Guilherme Dal Bianco

Aprovado em 12/12/2025.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
GUILHERME DAL BIANCO
Data: 17/12/2025 08:21:51-0300
Verifique em https://validar.iti.gov.br

Guilherme Dal Bianco – UFFS

gov.br Documento assinado digitalmente
GEOMAR ANDRE SCHREINER
Data: 18/12/2025 09:42:47-0300
Verifique em https://validar.iti.gov.br

Geomar André Schreiner – UFFS

gov.br Documento assinado digitalmente
SAMUEL DA SILVA FEITOSA
Data: 18/12/2025 10:32:47-0300
Verifique em https://validar.iti.gov.br

Samuel Da Silva Feitosa – UFFS

Validação automática de respostas técnicas automotivas utilizando modelos de linguagem e recuperação de informação

Rodrigo Ediel Vogt [**Universidade Federal da Fronteira Sul** | ***rodrigo.vgt@gmail.com***]
Guilherme Dal Bianco [**Universidade Federal da Fronteira Sul** | ***guilherme.dalbianco@uffs.edu.br***]

Resumo

Abstract Advances in large language models have enabled the automation of complex tasks involving the evaluation of technical content. This study investigates the use of Large Language Models to validate the coherence of Question–Answer (QA) pairs extracted from an automotive forum, comparing three approaches: zero-shot classification, retrieval based on semantically similar questions (few-shot), and retrieval based on excerpts from technical manuals.
Experimental results indicate a progressive improvement in performance as more structured context is introduced. While the zero-shot approach yields F1-scores below 50%, the use of similar questions as contextual examples provides moderate gains. The best results are achieved with manual-based retrieval, which increases the F1-score to over 70%, highlighting the importance of authoritative technical documentation for reliable automatic answer validation.

Resumo

O avanço dos modelos de linguagem de grande porte tem viabilizado a automação de tarefas complexas de avaliação de conteúdo técnico. Este trabalho investiga o uso desses modelos para validar a coerência entre pares de pergunta e resposta extraídos de um fórum automotivo, comparando três abordagens: classificação zero-shot, recuperação baseada em perguntas semanticamente semelhantes (few-shot) e recuperação a partir de trechos de manuais técnicos. Os resultados experimentais evidenciam uma melhoria progressiva no desempenho à medida que contextos mais estruturados são incorporados. Enquanto a abordagem zero-shot apresenta valores de F1 inferiores a 50%, o uso de perguntas semelhantes como exemplos contextuais promove ganhos moderados. O melhor desempenho é obtido com a recuperação baseada em manuais, que eleva o F1-score para valores superiores a 70%, ressaltando a relevância da documentação técnica oficial para a validação automática de respostas.

Palavras-chave: Modelos de Linguagem; GPT; Gemini; RAG; Validação de Respostas; Domínio Automotivo.

1 Introdução

O acesso à informação técnica tem se tornado um elemento central na resolução de problemas em diversas áreas, especialmente em contextos em que o conhecimento especializado é necessário para diagnósticos e para a tomada de decisão. Fóruns online cumprem papel relevante nesse cenário ao concentrar experiências práticas e soluções compartilhadas por profissionais e entusiastas, embora apresentem variações significativas na qualidade das respostas. Essa variabilidade torna desafiadora a identificação automática de conteúdo realmente útil.

Modelos de linguagem de grande porte, impulsionados por avanços recentes em aprendizado profundo, demonstraram habilidades notáveis na compreensão e na geração de texto, conforme demonstrado por Brown et al. [2020]. No entanto, mesmo modelos avançados podem apresentar limitações em tarefas que exigem conhecimento factual rigoroso [Ji et al., 2023]. Para mitigar essas limitações, abordagens que combinam modelos generativos com mecanismos de recuperação de informação, como o RAG, têm demonstrado ganhos substanciais ao incorporar evidências externas, conforme apresentado por Lewis et al. [2020].

O domínio automotivo é um exemplo de área em que esses desafios se intensificam. Fóruns especializados, como a "Oficina Brasil", concentram um grande volume de dúvidas e diagnósticos, mas dependem fortemente de contribuições humanas heterogêneas, em termos de precisão e profundidade. Trabalhos anteriores na documentação automotiva mostram que o uso de LLMs aliados a bases técnicas estruturadas pode melhorar significativamente a qualidade das respostas como visto em Medeiros et al. [2023].

Este trabalho investiga métodos automáticos para validar respostas técnicas automotivas utilizando modelos de linguagem e estratégias de recuperação de contexto. Para isso, foi construída uma base contendo pares de pergunta e resposta provenientes de um fórum real, complementada por trechos de manuais automotivos oficiais. Três abordagens foram avaliadas: classificação *zero-shot*, *few-shot learning* e recuperação baseada em documentação técnica. Os resultados são analisados com base em métricas de desempenho, fornecendo evidências sobre a eficácia e as limitações de cada método no contexto estudado.

Nas próximas seções, apresenta-se a fundamentação teórica que embasa o estudo, seguida da metodologia adotada para a construção da base, a execução dos experimentos e a avaliação dos modelos. Em seguida, são discutidos os resultados obtidos e, por fim, apresentam-se as conclusões e as sugestões para trabalhos futuros.

2 Fundamentação Teórica

Esta seção apresenta os principais conceitos teóricos que embasam o desenvolvimento deste trabalho. Inicialmente, são discutidos os fundamentos dos modelos de linguagem de grande porte, destacando suas capacidades e limitações em tarefas de interpretação e validação de conteúdo técnico. Em seguida, são abordados os paradigmas de *zero-shot* e *few-shot learning*, que descrevem diferentes regimes de inferência baseados na ausência ou na presença de exemplos contextuais. Posteriormente, são apresentados os métodos de Recuperação de Informação e a abordagem de *Retrieval-Augmented Generation* (RAG), que integram modelos de linguagem a fontes externas de conhecimento. Por fim, são discutidos os conceitos de bases vetoriais e *embbedings*, fundamentais para a recuperação semântica dos contextos utilizados nos experimentos realizados.

2.1 Modelos de Linguagem (LLMs)

Modelos de linguagem de grande porte *(Large Language Models, LLMs)* representam um marco no processamento de linguagem natural. Baseados na arquitetura *Transformer* proposta inicialmente por Vaswani et al. [2017], esses modelos conseguem capturar dependências de longo alcance e produzir textos coerentes em diferentes contextos. Exemplos recentes incluem o GPT, treinado pela OpenAI [Brown et al., 2020, OpenAI, 2024], e o Gemini, desenvolvido pela Google DeepMind [DeepMind, 2023]. Esses modelos demonstram a capacidade de compreender relações semânticas e realizar tarefas complexas, como classificação e validação de conteúdos técnicos.

2.2 Zero-shot e Few-shot Learning

Na abordagem *zero-shot*, o modelo é solicitado a resolver uma tarefa sem o fornecimento de exemplos explícitos, baseando-se exclusivamente no conhecimento adquirido durante o pré-treinamento. Esse paradigma foi formalizado por Brown et al. [2020], que demonstraram a capacidade de modelos de linguagem de grande porte em generalizar para novas tarefas apenas a partir de instruções textuais. Embora simples e eficiente, o zero-shot tende a apresentar limitações em domínios técnicos, nos quais a validação correta de respostas depende de conhecimento específico e contextualizado.

O *few-shot learning*, por sua vez, consiste no fornecimento de um pequeno conjunto de exemplos representativos da tarefa como parte do contexto de entrada do modelo. Esses exemplos funcionam como demonstrações implícitas, orientando o comportamento do modelo durante a inferência [Brown et al., 2020]. Diferentemente de abordagens supervisionadas tradicionais, o *few-shot* não requer ajuste de parâmetros, sendo realizado inteiramente no nível do *prompt*.

Em cenários práticos, os exemplos utilizados no *few-shot* podem ser selecionados dinamicamente a partir de bases de dados, utilizando medidas de similaridade semântica. Essa estratégia permite que o modelo seja exposto a casos relevantes e contextualizados, mesmo na ausência de rótulos explícitos ou curadoria manual. No contexto deste trabalho, o *few-shot* é implementado por meio da recuperação de perguntas semanticamente semelhantes, que servem como exemplos contextuais para auxiliar na validação da coerência entre perguntas e respostas técnicas.

2.3 Recuperação de Informação e *Retrieval-Augmented Generation(RAG)*

Embora o *few-shot learning* permita orientar o comportamento dos modelos de linguagem por meio de exemplos contextuais, essa abordagem ainda depende da qualidade e da completude das informações presentes nos próprios dados recuperados. Em domínios técnicos, como o automotivo, respostas humanas podem conter ambiguidades, omissões ou interpretações incorretas, o que limita a confiabilidade do contexto fornecido ao modelo.

Como alternativa, métodos baseados em Recuperação de Informação (*Information Retrieval*) têm sido empregados para complementar o conhecimento interno dos modelos com fontes externas mais estruturadas. A abordagem conhecida como *Retrieval-Augmented Generation* (RAG) combina mecanismos de busca semântica com modelos de linguagem, permitindo que documentos relevantes sejam recuperados e incorporados ao contexto de entrada antes da geração ou avaliação de uma resposta, conforme proposto por Lewis et al. [2020].

Nessa arquitetura, textos provenientes de fontes externas são convertidos em representações vetoriais, possibilitando a recuperação eficiente de trechos semanticamente similares por meio de medidas de proximidade em espaços de alta dimensionalidade. Esses trechos são então utilizados como evidência contextual, reduzindo a dependência exclusiva do conhecimento paramétrico do modelo e mitigando a ocorrência de erros factuais e alucinações, conforme observado em trabalhos anteriores [Guu et al., 2020].

No contexto deste trabalho, a estratégia de RAG é aplicada à validação de respostas técnicas automotivas por meio da recuperação de trechos de manuais oficiais de fabricantes. Diferentemente do few-shot baseado em exemplos de fóruns, essa abordagem fornece ao modelo informações normativas e tecnicamente validadas, permitindo avaliar a coerência das respostas com base em documentação especializada e confiável.

2.4 Bases vetoriais

A representação vetorial permite comparar textos com base na proximidade semântica, e medidas como a similaridade do cosseno são amplamente utilizadas em sistemas de busca. Para tornar esse processo eficiente em grande escala, bases vetoriais empregam estruturas de busca por vizinhos mais próximos *Approximate Nearest Neighbors* — ANN), possibilitando operações rápidas mesmo em coleções extensas, conforme discutido por Johnson et al. [2019].

Esse tipo de indexação é fundamental em métodos de recuperação densa (*dense retrieval*), nos quais os documentos são convertidos em *embeddings* capazes de capturar relações semânticas mais profundas, superando as limitações das abordagens lexicais tradicionais. Em domínios técnicos — como o automotivo — isso é particularmente útil, pois sintomas

e falhas podem ser descritos com vocabulário variável, mas ainda assim mantêm forte similaridade semântica, conforme observado por Guo et al. [2020].

Ferramentas como o ChromaDB implementam essas estruturas para o armazenamento e a consulta eficientes de *embeddings*, permitindo a recuperação de itens *top_k* relevantes para a composição de contexto. Quando combinadas às técnicas de *chunking* com sobreposição, essas bases viabilizam pipelines de Recuperação Aumentada por Geração (*Retrieval-Augmented Generation*) ao fornecerem ao modelo trechos coerentes e semanticamente pertinentes, alinhando-se ao fluxo proposto por Lewis et al. [2020].

3 Trabalhos Relacionados

O uso de Modelos de Linguagem em tarefas de análise e validação de conteúdo tem avançado significativamente, com diversos estudos demonstrando sua capacidade de interpretar contextos e capturar relações semânticas complexas, como visto em Zhou et al. [2023]. Pesquisas sobre verificação factual [Thorne et al., 2018] e classificação automática mostram que esses modelos podem auxiliar na avaliação de respostas em diferentes domínios técnicos, como discutido em Ko et al. [2023]. Além disso, Brown et al. [2020] analisam a capacidade dos modelos de resolver tarefas sem exemplos adicionais, enquanto Lewis et al. [2020] exploram a integração entre modelos generativos e mecanismos de recuperação semântica, mostrando que essa combinação pode reduzir alucinações e aumentar a precisão das respostas.

O uso de *embeddings* e bases vetoriais tem sido amplamente explorado em sistemas que dependem da recuperação de trechos específicos para apoiar as respostas de modelos de linguagem. Essa abordagem aparece em diferentes domínios, como o médico, em Singh et al. [2023], e o jurídico, em Chalkidis et al. [2022], além de métodos gerais de recuperação densa baseados em vizinhança semântica, como discutido em Khandelwal et al. [2020]. No contexto automotivo, estudos como o de Medeiros et al. [2023] demonstram que a combinação entre modelos de linguagem e recuperação de informações técnicas pode melhorar a precisão das respostas e apoiar atividades de diagnóstico e suporte.

Apesar dos avanços no uso de Modelos de Linguagem, ainda há poucos estudos voltados especificamente à validação de respostas produzidas por humanos em fóruns técnicos, cuja qualidade heterogênea apresenta desafios adicionais para modelos automáticos. Nesse contexto, o presente trabalho contribui ao comparar abordagens com diferentes níveis de contextualização e ao avaliar a capacidade dos modelos de distinguir respostas corretas e incorretas em um domínio especializado.

Com base na análise dos trabalhos apresentados, a Tabela 1 sintetiza as principais diferenças entre os trabalhos relacionados e a abordagem proposta neste estudo. Observa-se que trabalhos fundamentais da literatura estabelecem as bases conceituais do uso de modelos de linguagem em regimes *zero-shot* e *few-shot*, bem como a integração com mecanismos de recuperação de informação. No entanto, essas abordagens são, em geral, avaliadas em domínios genéricos e voltadas à geração de respostas, sem foco específico na validação automática de conteúdo produzido por humanos.

Por outro lado, estudos recentes no domínio automotivo exploram a utilização de documentação técnica oficial para suporte à resolução de consultas, mas não realizam uma análise comparativa entre diferentes níveis de contextualização nem consideram explicitamente cenários *zero-shot* e few-shot. Nesse contexto, o presente trabalho diferencia-se ao integrar e comparar sistematicamente abordagens *zero-shot*, *few-shot* e RAG baseadas em manuais automotivos, aplicadas à validação de respostas técnicas extraídas de um fórum real, evidenciando o impacto progressivo da recuperação de contexto estruturado sobre o desempenho dos modelos.

4 Configuração Experimental

Antes da apresentação detalhada da metodologia, esta seção descreve as abordagens avaliadas no estudo. O objetivo é fornecer uma visão geral dos três métodos comparados, destacando o tipo de informação utilizada por cada um e a forma como contribuem para a validação de perguntas-resposta.

4.1 Zero-shot

Na abordagem *zero-shot*, o modelo recebe apenas a pergunta e a resposta a serem avaliadas, sem contexto adicional. O objetivo é verificar a capacidade dos modelos de linguagem de julgar a coerência entre os dois elementos exclusivamente com base em seu conhecimento interno.

4.2 Few-Shot baseado em Perguntas Semelhantes

Nesta abordagem, são recuperadas perguntas semelhantes presentes na base de dados. Esses exemplos servem como contexto adicional, permitindo que o modelo compare a resposta analisada a soluções reais para problemas semelhantes. O método explora a capacidade dos *embeddings* de identificar proximidade semântica entre consultas.

4.3 RAG baseado em Manuais Automotivos

Aqui, o modelo recebe trechos de manuais técnicos como contexto. Esses textos fornecem informações oficiais de fabricantes sobre funcionamento e diagnóstico automotivo. O objetivo é avaliar se o modelo consegue utilizar documentação técnica para validar se a resposta está de acordo com orientações e padrões presentes nos manuais.

A Figura 1 apresenta um resumo visual das três abordagens, destacando o tipo de entrada utilizado por cada método.

5 Metodologia

Nesta seção, será apresentada a metodologia adotada para a coleta, preparação e avaliação dos dados utilizados no estudo. São descritos o processo de extração de perguntas e respostas do fórum automotivo, os critérios de curadoria e a construção das bases verdadeira e falsa, a segmentação dos manuais técnicos e sua indexação em uma base vetorial, bem

Trabalho	Domínio	LLMs	Zero-shot	Few-shot	RAG	Validação de respostas humanas	Fonte de contexto
Brown et al. (2020)	Geral	✓	✓	✓	×	×	Nenhuma
Lewis et al. (2020)	Geral	✓	×	×	✓	×	Documentos externos genéricos
Guu et al. (2020)	Geral	✓	×	×	✓	×	Base textual não estruturada
Medeiros et al. (2023)	Automotivo	✓	×	×	✓	×	Manuais automotivos oficiais
Este trabalho	**Automotivo**	✓	✓	✓	✓	✓	**Fórum técnico + Manuais oficiais**

Tabela 1. Comparação entre trabalhos relacionados e a abordagem proposta

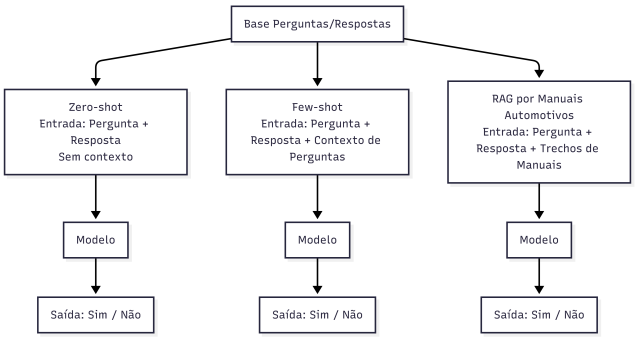

Figura 1. Visão geral das três abordagens avaliadas: zero-shot, few-shot e RAG baseado em manuais automotivos.

como as três abordagens de inferência avaliadas. Também são detalhados os modelos utilizados e os procedimentos de recuperação de contexto. Esse conjunto de etapas constitui o pipeline experimental que fundamenta os resultados discutidos posteriormente.

O código pode ser acessado no repositório: `https://github.com/RodrigoVgt/tcc_II`

5.1 Coleta de Dados no Fórum

A escolha de um fórum técnico como fonte primária de dados se justifica pelo papel central que esse tipo de plataforma desempenha na prática cotidiana da área automotiva. Fóruns reúnem relatos reais de problemas, diagnósticos realizados por profissionais e por entusiastas, além de discussões que refletem o processo prático de solução de falhas em veículos. Diferentemente de bases acadêmicas ou de documentações oficiais, o conteúdo gerado por usuários permite observar como o conhecimento técnico circula em situações concretas, incluindo erros comuns, soluções consolidadas e interpretações equivocadas.

Além disso, muitos fóruns adotam mecanismos internos de validação, como o marcador, neste caso, representado por um troféu, que indica a resposta considerada mais útil ou correta pela comunidade. Esse mecanismo oferece um sinal preliminar de qualidade, permitindo assumir que grande parte dessas respostas representa soluções plausíveis no contexto automotivo. Dessa forma, o uso do fórum fornece um conjunto realista, variado e representativo para a construção da base de validação deste estudo.

A obtenção dos dados utilizados neste trabalho foi realizada por meio de um processo de *scraping* aplicado ao fórum *Oficina Brasil*, reconhecido como um dos principais repositórios nacionais de dúvidas, diagnósticos e discussões técnicas na área automotiva. Conforme ilustrado na Figura 2, a plataforma organiza suas interações em formato de perguntas e respostas, permitindo que os próprios usuários indiquem a "melhor resposta" para cada tópico. Esse mecanismo funciona como uma forma de curadoria comunitária, sugerindo

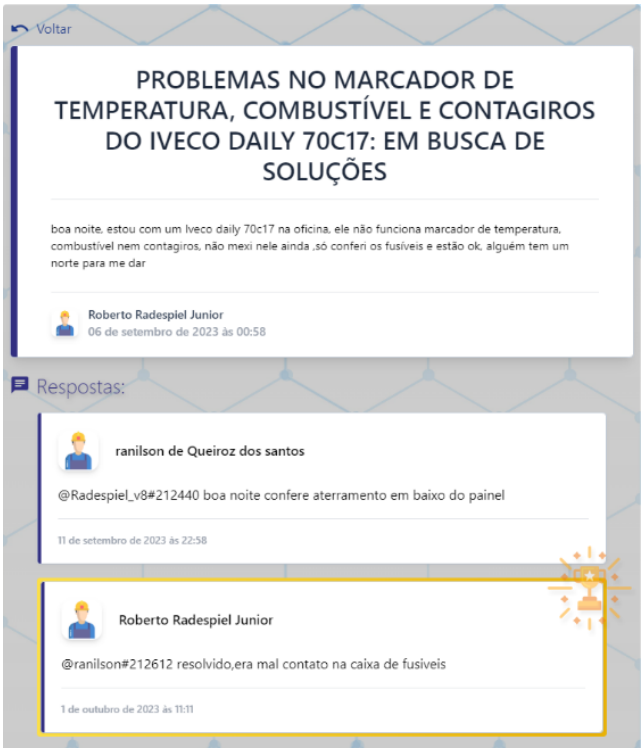

Figura 2. Representação do fórum a partir do qual foram extraídas as perguntas e as respostas. A resposta em amarelo indica a melhor resposta escolhida.

que a resposta marcada é, em geral, a que melhor solucionou o problema apresentado.

A coleta dos dados foi implementada utilizando a biblioteca *Puppeteer* em ambiente *Node.js*, por meio de um *script* independente do restante da *pipeline* experimental. Essa abordagem possibilitou extrair de maneira estruturada o conteúdo textual das postagens, ao mesmo tempo em que contornou limitações decorrentes do carregamento dinâmico do site, garantindo a captura confiável das informações necessárias para a construção da base de dados.

Foram coletados todos os conjuntos completos de pergunta e resposta disponíveis nas discussões, sem filtragem prévia, incluindo metadados essenciais para análises posteriores: identificador do usuário, data e hora da publicação, categoria da postagem e o marcador *best_answer*, utilizado pela própria plataforma para indicar a resposta considerada a melhor solução pelos participantes. A extração do conteúdo textual foi realizada diretamente do corpo das postagens, contemplando tanto a pergunta quanto todas as respostas associadas ao tópico.

O fórum oferece um sistema de categorização interno, incluindo áreas voltadas a tipos distintos de veículos e sistemas automotivos. Para este estudo, foi aplicada a filtragem nativa da plataforma, limitando a coleta à categoria "Leve", que abrange veículos de passeio e corresponde à maior parte das discussões relacionadas a diagnósticos comuns. Essa se-

leção inicial permitiu concentrar o escopo da análise em um domínio consistente, reduzindo ruído e evitando tópicos não relevantes, como discussões sobre veículos pesados ou máquinas agrícolas.

5.2 Filtragem e Curadoria dos Dados

Após a coleta inicial, foi realizada uma etapa de filtragem destinada a remover conteúdos que não atendiam aos requisitos para a validação automática da coerência entre pergunta e resposta. Embora todas as interações tenham sido extraídas integralmente do fórum, apenas parte delas apresentava estrutura e autonomia suficientes para compor a base utilizada nos experimentos.

Para identificar respostas irrelevantes, foi desenvolvido um conjunto de heurísticas baseadas em padrões textuais (*regex*), capazes de detectar mensagens cujo propósito não fosse oferecer uma solução técnica à pergunta. Entre esses padrões, destaca-se a eliminação de respostas de agradecimento, frequentemente compostas por expressões como ”obrigado”, ”valeu”ou variações semelhantes. Além disso, foram removidas respostas que dependiam explicitamente de outras mensagens para fazer sentido, identificadas por expressões como *”como disse o colega”*, que indicam dependência contextual entre respostas distintas.

Outro critério de exclusão envolveu respostas compostas exclusivamente por imagens ou que as contivessem como elemento central. Esse tipo de conteúdo não é adequado para análise baseada em modelos de linguagem, e sua detecção foi facilitada pelo atributo `has_image` presente nos dados coletados, permitindo descartar automaticamente mensagens que não contavam com conteúdo textual suficiente.

Ao final desse processo de curadoria, chegou-se a um conjunto final composto por pares pergunta/resposta considerados válidos e sem inconsistências estruturais, totalizando 492 instâncias utilizadas como base para as etapas subsequentes do estudo.

5.3 Construção das Bases de Dados

A partir do conjunto final, composto por 492 pares pergunta/resposta válidos, foi construída uma base com respostas falsas sinteticamente geradas, em igual quantidade. A geração de respostas falsas adotou uma estratégia híbrida, combinando aleatoriedade controlada e verificações semânticas. Em um primeiro momento, respostas verdadeiras foram associadas a perguntas aleatórias distintas, gerando pares artificialmente incorretos. No entanto, para evitar a formação de respostas completamente desconexas ou absurdas, foram aplicadas verificações intermediárias baseadas em similaridade semântica, assegurando que as respostas falsas permanecessem linguisticamente plausíveis, embora não corretas em relação à pergunta apresentada. Esse cuidado foi essencial para tornar a tarefa de validação mais desafiadora, reduzindo o risco de que os modelos identificassem pares inválidos apenas por incongruências linguísticas superficiais.

O resultado final foi um conjunto perfeitamente balanceado, contendo 492 respostas verdadeiras e 492 respostas falsas, garantindo simetria entre as classes e evitando viés nos processos de treinamento e avaliação dos métodos aplicados.

Para os testes, foram criados três grupos distintos a partir da combinação entre as bases verdadeiras e falsas. Também foram separadas duas bases a partir das misturas geradas anteriormente, cada uma com 492 respostas, distribuídas de forma igual.

Com o objetivo de reduzir o impacto de variações decorrentes da seleção aleatória das instâncias, o conjunto final foi particionado em três grupos independentes de teste, aqui tratados como *folds*. Diferentemente de esquemas tradicionais de validação cruzada empregados em cenários de treinamento supervisionado, não há, neste trabalho, uma fase de treinamento dos modelos. Cada *fold* é utilizado exclusivamente para avaliação, permitindo observar a estabilidade do desempenho dos métodos em diferentes subconjuntos balanceados da base.

Os resultados reportados ao longo do trabalho correspondem à média das métricas obtidas nesses três *folds*, acompanhadas de seus respectivos intervalos de confiança. Essa estratégia permite mitigar o efeito de flutuações amostrais e fornece uma estimativa mais robusta do desempenho dos modelos, especialmente relevante em tarefas de validação automática em domínios técnicos, nas quais pequenas variações no conjunto de teste podem influenciar significativamente os resultados observados.

Cada *fold* foi formado por amostragem aleatória simples, sem reposição, selecionando-se 246 pares da base verdadeira e 246 pares da base falsa. Esse procedimento garantiu que cada subconjunto fosse perfeitamente balanceado, contendo 50% de respostas verdadeiras e 50% de respostas falsas, condição necessária para uma comparação equitativa entre abordagens e modelos. O processo foi repetido integralmente para a criação dos três grupos, reiniciando sempre a base original antes de cada nova seleção. Dessa forma, os grupos são independentes entre si, embora todos sejam derivados da mesma base completa de 984 instâncias. A utilização de múltiplos grupos permite avaliar a estabilidade dos métodos propostos e mitigar o efeito de ”sorte amostral”, aumentando a confiabilidade das análises estatísticas e dos resultados apresentados.

5.4 *Tokenização* e Separação dos Manuais

Para a etapa de avaliação baseada em trechos de manuais automotivos, tornou-se necessário selecionar quais documentos serviriam de fonte de contexto técnico. Essa seleção não foi arbitrária, uma vez que diferentes modelos de veículos apresentam variações significativas em seus sistemas mecânicos e elétricos, o que torna irrelevante o uso de manuais pouco representados no conjunto de perguntas. Com esse objetivo, foi realizada uma análise de frequência sobre todas as perguntas da base verdadeira, identificando as marcas e modelos mais mencionados pelos usuários. Um *script* de reconhecimento de padrões foi empregado para extrair e padronizar essas referências, permitindo contabilizar de forma consistente os modelos citados. A partir desse levantamento, foram selecionados apenas os modelos que apresentavam pelo menos três ocorrências na base, de modo a priorizar manuais com maior relevância prática e reduzir o impacto de casos isolados. Esse critério resultou na escolha de sete manuais, representando os veículos mais frequentemente discutidos no

conjunto analisado. Uma correlação de marcas, modelos e a quantidade de aparições pode ser encontrada na Tabela 2

Marca	Modelo	Ocorrências
Volkswagen	Gol	11
Ford	Ka	9
Fiat	Uno	6
Chevrolet	Onix	6
Ford	Ecosport	5
Ford	Fiesta	4
Ford	Fusion	3

Tabela 2. Frequência de modelos automotivos presentes na base de perguntas.

Após a definição dos modelos, foi necessária a obtenção dos respectivos manuais técnicos. Para isso, optou-se pela plataforma Manual do Proprietário (`https://www.manualcarro.com.br/`), que disponibiliza manuais oficiais organizados por marca, modelo e ano de fabricação. A escolha dessa base se deve à sua abrangência, à atualização contínua e à facilidade de acesso aos documentos originais, fatores essenciais para garantir consistência e confiabilidade na extração dos trechos utilizados como contexto nos métodos de recuperação de informação.

Após a conversão dos manuais para texto contínuo, foi realizada a segmentação do conteúdo em unidades menores com o objetivo de possibilitar a geração de *embeddings* e posterior indexação no mecanismo de recuperação. A segmentação adotou uma abordagem baseada em *tokens*, permitindo maior controle sobre o tamanho semântico dos trechos e respeitando os limites operacionais dos modelos utilizados.

A tokenização foi realizada utilizando a biblioteca `tiktoken`, empregando o codificador `cl100k_base`, compatível com modelos da família GPT que adotam o mesmo esquema de tokenização. Cada manual foi convertido em uma sequência de *tokens* e, em seguida, dividido em *chunks* de 500 *tokens*, valor escolhido por corresponder aproximadamente ao conteúdo típico de uma página de manual automotivo. Para preservar a continuidade semântica entre as janelas de texto, foi aplicada uma sobreposição de 50 *tokens* entre segmentos consecutivos.

Esse procedimento resultou em um conjunto estruturado de trechos textuais padronizados, adequados à geração de *embeddings* e à recuperação posterior. A combinação entre tamanho fixo, sobreposição controlada e tokenização consistente garantiu que informações importantes não fossem fragmentadas de forma inadequada, mantendo a coerência necessária para as consultas contextuais realizadas pelos modelos avaliados.

5.5 Construção da Base Vetorial

Para possibilitar a recuperação eficiente de trechos relevantes, tanto de perguntas quanto de manuais automotivos, foi construída uma base vetorial utilizando o ChromaDB. A geração dos *embeddings* utilizados neste trabalho foi realizada de forma padronizada em todas as etapas, empregando o modelo *text-embedding-3-small*, disponibilizado pela OpenAI. A escolha desse modelo se baseou em três fatores principais: custo reduzido, desempenho consistente em *benchmarks* públicos e adequação à tarefa de recuperação semântica em domínios amplos.

Em avaliações independentes, modelos de *embedding* recentes da OpenAI demonstraram desempenho competitivo em tarefas de busca semântica, classificação textual e similaridade, apresentando resultados expressivos em conjuntos de avaliação como o MTEB (*Massive Text Embedding Benchmark*) [Muennighoff et al., 2022]. Estudos recentes também indicam que *embeddings* densos obtidos por técnicas de aprendizado contrastivo tendem a apresentar maior robustez e generalização, especialmente em cenários com diversidade lexical e respostas heterogêneas [Reimers and Gurevych, 2019].

O *text-embedding-3-small* apresenta custo operacional significativamente inferior ao de outras alternativas disponíveis no mercado, o que foi determinante para sua adoção, considerando o volume substancial de texto a ser processado, especialmente durante a tokenização e vetorização dos manuais automotivos. A ausência de limitações rígidas de requisições simultâneas também permitiu a construção eficiente de toda a base vetorial, garantindo uniformidade e estabilidade nos resultados das consultas posteriores.

Como as perguntas e respostas pertenciam a três grupos independentes, cada grupo recebeu sua própria coleção no ChromaDB, evitando interferências entre subconjuntos distintos. Essa separação garante que a recuperação de similaridade dentro de cada grupo seja conduzida exclusivamente com base nos elementos que o compõem, reduzindo qualquer risco de viés derivado de instâncias externas. Por outro lado, os manuais automotivos, por se tratarem de uma fonte estática e comum a todos os experimentos, foram armazenados em uma única coleção dedicada, independentemente das coleções de perguntas.

Os dados armazenados variaram conforme a fonte. Para os manuais, foram indexados apenas os trechos segmentados em *chunks*, enquanto, para as perguntas, foram armazenados também metadados adicionais, incluindo a melhor resposta associada e o rótulo de validade do par. Esses metadados foram inseridos para permitir sua utilização como contexto direto na construção dos *prompts* nos métodos de recuperação.

A recuperação dos elementos mais semelhantes foi realizada diretamente via função `query` de cada coleção, permitindo obter os trechos com maior similaridade para diferentes valores de *top_k*. Essa consulta foi aplicada sistematicamente para *top_k* 1, 3 e 5, tanto para coleções de perguntas quanto para a coleção de manuais, assegurando comparabilidade entre todos os cenários avaliados. Todas as etapas subsequentes do experimento utilizaram exclusivamente essas mesmas estruturas vetoriais, garantindo reprodutibilidade e consistência ao longo de todas as execuções.

5.6 Abordagens Avaliadas

Foram avaliadas três abordagens distintas para a validação automática da relação pergunta/resposta: (i) classificação direta em modo *zero-shot*, (ii) *few-shot* com base em perguntas semelhantes e (iii) recuperação com base em trechos de manuais automotivos. Todas as abordagens utilizaram o mesmo

formato de *prompt* e retornos binários, garantindo comparabilidade entre os cenários.

5.6.1 Zero-shot

Na abordagem *zero-shot*, o *prompt* consistiu unicamente na pergunta, na resposta a ser avaliada e em uma instrução explícita solicitando que o modelo informasse, de forma binária, se a resposta era válida ou não. A saída esperada era sempre "Sim"ou "Não", sem justificativas adicionais. Esse formato foi padronizado para todos os modelos avaliados e executado com temperatura fixa, de modo a eliminar variações decorrentes de aleatoriedade na geração.

A cada consulta, o resultado produzido pelo modelo era comparado com o rótulo verdadeiro do par, armazenando-se imediatamente se o modelo havia acertado ou não. Esse procedimento simplificou a etapa posterior de análise e cálculo de métricas.

5.6.2 *few-shot* baseado em Perguntas Semelhantes

Na abordagem de *few-shot* baseada em perguntas semelhantes, foram utilizados os *embeddings* armazenados no *ChromaDB* para identificar, para cada pergunta avaliada, os itens semanticamente mais próximos pertencentes ao mesmo grupo de teste. Para cada valor de *top_k* 1, 3 e 5, os trechos mais similares foram recuperados e organizados como um bloco de contexto adicional.

O *prompt* final apresentava ao modelo uma descrição do contexto extraído, seguida da pergunta e da resposta a serem avaliadas. A instrução explicitava a tarefa de verificar se a resposta era válida à luz do contexto oferecido, com a exigência de que o modelo respondesse exclusivamente "Sim"ou "Não". Assim como no método *zero-shot*, a temperatura permaneceu fixa em todas as execuções. O resultado binário era então registrado e comparado ao rótulo verdadeiro correspondente.

5.6.3 RAG baseado em Manuais Automotivos

Na abordagem baseada em manuais automotivos, os trechos recuperados correspondiam aos *chunks* textuais previamente extraídos e segmentados a partir dos manuais técnicos dos fabricantes. Para cada consulta, foram recuperados os segmentos mais relevantes, segundo a similaridade vetorial, considerando novamente os valores de *top_k* 1, 3 e 5. Todos os trechos recuperados eram agrupados em um único bloco contínuo de texto, mantendo a ordem retornada pelo mecanismo de busca.

O *prompt* empregado nessa abordagem apresentava o contexto como "Contexto extraído do manual", seguido pela pergunta e pela resposta analisada. Assim como nos demais métodos, o modelo era instruído a atuar como um avaliador técnico automotivo, limitando sua saída a uma das duas respostas possíveis: "Sim", se a resposta fosse válida para a pergunta apresentada, ou "Não", caso contrário. Nenhum limite adicional de *tokens* foi necessário, pois os valores utilizados estavam confortavelmente abaixo dos limites suportados pelas APIs.

A saída do modelo era novamente armazenada para comparação direta com o rótulo verdadeiro, permitindo a avaliação detalhada do impacto do contexto técnico dos manuais nos diferentes cenários testados.

5.6.4 Configuração dos Modelos e Parâmetros

Todos os métodos descritos anteriormente foram executados com dois modelos de linguagem de grande porte: o *GPT-4o-mini*, disponibilizado pela OpenAI, e o *Gemini-2.5-flash-lite*, disponibilizado pela Google. Ambos os modelos foram empregados de forma consistente em todos os cenários avaliados, permitindo comparações diretas entre as arquiteturas.

A temperatura foi mantida constante em todas as chamadas, assegurando determinismo e evitando variações decorrentes de amostragem estocástica. Os modelos foram instruídos a responder exclusivamente com uma das duas saídas possíveis: "Sim", se a resposta analisada fosse considerada válida para a pergunta apresentada, ou "Não", caso contrário. Nenhuma justificativa textual foi permitida.

O mesmo formato de *prompt* foi adotado em todos os métodos, variando apenas a presença ou ausência de blocos de contexto, provenientes de perguntas semelhantes ou de trechos de manuais automotivos. Após cada inferência, o resultado binário produzido pelo modelo era imediatamente comparado ao rótulo verdadeiro correspondente e armazenado, facilitando o cálculo das métricas adotadas ao longo do estudo.

5.7 Métricas de Avaliação

A avaliação do desempenho dos métodos propostos foi realizada por meio de métricas derivadas da matriz de confusão, composta por quatro componentes fundamentais: verdadeiros positivos (TP), falsos positivos (FP), verdadeiros negativos (TN) e falsos negativos (FN).

A acurácia foi utilizada como medida geral de desempenho, representando a proporção total de classificações corretas entre todas as instâncias avaliadas. A precisão quantificou a proporção de predições positivas que eram, de fato, positivas, refletindo a capacidade do modelo de evitar classificações incorretas como válidas. O *recall*, ou sensibilidade, mensurou a capacidade do modelo de identificar corretamente as respostas válidas presentes na base. Já o F1-score, definido como a média harmônica entre precisão e *recall*, forneceu uma medida equilibrada entre ambos, sendo especialmente relevante em cenários nos quais erros falso-positivos e falso-negativos possuem impacto semelhante. As definições dessas métricas seguem a formulação clássica apresentada em Manning et al. [2008].

Além das métricas pontuais, foram calculados intervalos de confiança para quantificar a incerteza associada aos resultados. Para métricas expressas como proporções diretas, tais como acurácia, precisão e *recall*, foi aplicado o intervalo de confiança de Wilson, que apresenta maior estabilidade do que o intervalo normal-padrão em amostras de tamanho moderado.

5.8 Pipeline Geral dos Experimentos

A execução dos experimentos seguiu uma sequência organizada de etapas, responsável por integrar a recuperação de contexto, a avaliação pelos modelos e o cálculo das métricas finais. Inicialmente, cada um dos três grupos de teste independentes era carregado individualmente. Em seguida, o método de avaliação selecionado, *zero-shot* foi executado sem o fornecimento de contextos. Para os experimentos com contexto, tanto de perguntas quanto de respostas, foi executada a recuperação do contexto baseando-se no número de top_k da rodada em questão .Após a definição do contexto, o *prompt* correspondente era construído e enviado ao modelo selecionado, mantendo-se os mesmos parâmetros em todas as execuções.

Para todos os testes, a resposta binária produzida (”Sim”ou ”Não”) era imediatamente registrada e comparada ao rótulo verdadeiro associado ao par pergunta/resposta.

Esse pipeline unificado assegurou consistência entre os métodos, facilitando a comparação entre modelos e entre diferentes configurações de recuperação de contexto.

5.9 Considerações Finais

O conjunto de procedimentos descrito nesta seção estabelece uma estrutura experimental completa e reprodutível para a avaliação dos modelos de linguagem em diferentes cenários testados. Desde a coleta e curadoria dos dados, passando pela construção da base de dados e pela segmentação e indexação dos manuais, até a definição dos métodos avaliados e das métricas estatísticas empregadas, todas as etapas foram cuidadosamente projetadas para garantir consistência, controle experimental e comparabilidade entre as abordagens. Com esse pipeline consolidado, torna-se possível analisar, de forma sistemática, o impacto da recuperação de contexto e o comportamento dos modelos em múltiplos cenários de validação, cujos resultados são apresentados na próxima seção.

6 Resultados

Esta seção apresenta o desempenho dos modelos nos três cenários experimentais avaliados: *zero-shot*, *few-shot* de perguntas semelhantes e RAG-Manuais. Os valores reportados correspondem à média das métricas obtidas nos três *folds* independentes de teste, acompanhados de seus respectivos intervalos de confiança de 95%, separados por modelos e top_k contextos fornecidos a cada abordagem.

O desempenho inferior observado no cenário zero-shot reflete a ausência de contexto externo, conforme discutido na fundamentação teórica. Conforme apresentado na Tabela 3, o GPT obteve F1 médio de aproximadamente 45,85 % $\pm$ 5,02 pontos percentuais, evidenciando alta variabilidade entre os grupos avaliados. O Gemini apresentou desempenho similar, com F1 médio de 42,85 % $\pm$ 1,35 p.p., mostrando menor oscilação entre grupos, porém desempenho igualmente limitado quando comparado às abordagens que incorporam contexto externo. Esses resultados reforçam a limitação dos modelos ao dependerem exclusivamente de seu conhecimento interno, especialmente em um domínio técnico como o automotivo, em que a validade de uma resposta frequentemente exige referência a procedimentos específicos ou informações normativas.

Modelo	Precisão	Recall	F1
GPT	69.58 ± 7.95	33.73 ± 4.88	45.85 ± 5.02
Gemini	47.52 ± 1.47	39.02 ± 1.38	42.85 ± 1.35

Tabela 3. Desempenho no cenário Zero-shot

A introdução de contexto por meio do *few-shot* de perguntas semanticamente semelhantes elevou o desempenho de ambos os modelos, conforme apresentado na Tabela 4. No caso do GPT, o melhor resultado foi obtido com $k = 3$, com F1 médio de $54{,}08\% \pm 1{,}37$. Para $k = 1$ e $k = 5$, os valores foram, respectivamente, $49{,}91\% \pm 0{,}44$ e $51{,}12\% \pm 1{,}24$. O modelo Gemini apresentou comportamento semelhante, alcançando seu melhor desempenho também em $k = 3$, com F1 médio de $51{,}10\% \pm 0{,}80$. Para $k = 1$ e $k = 5$, os valores observados foram $46{,}66\% \pm 0{,}34$ e $47{,}51\% \pm 0{,}55$. Esses resultados demonstram que, embora o ganho em relação ao cenário *zero-shot* seja consistente, o impacto da quantidade de contextos recuperados apresenta variação moderada, sugerindo sensibilidade ao tipo e à relevância das perguntas similares retornadas pelo mecanismo de busca semântica.

O melhor desempenho foi obtido com o uso de trechos de manuais automotivos, conforme apresentado na Tabela 5. Esse tipo de contexto fornece ao modelo informações técnicas precisas e alinhadas à documentação oficial dos fabricantes, reduzindo ambiguidades e aprimorando a capacidade de verificar a validade das respostas. No caso do GPT, o F1 elevou-se de aproximadamente $70{,}64\%$ para $73{,}51\%$ quando o número de itens recuperados passou de $k = 1$ para $k = 3$, representando um ganho relativo de cerca de 4%. Ao aumentar para $k = 5$, o F1 manteve-se estável em $73{,}42\%$, configurando um ganho total próximo de 4% em relação ao cenário inicial.

O Gemini apresentou desempenho inferior ao GPT em todos os cenários, iniciando com F1 de $61{,}30\%$ para $k = 1$ e alcançando $65{,}11\%$ em $k = 3$, o que corresponde a um aumento relativo de aproximadamente 6%. Para $k = 5$, entretanto, o desempenho caiu para $63{,}43\%$, reduzindo o ganho acumulado para cerca de 3% em comparação com o cenário de $k = 1$.

Esses resultados evidenciam que a inclusão de documentação técnica é substancialmente mais eficaz do que o uso de exemplos extraídos do próprio fórum. Isto possivelmente se deve ao caráter normativo e detalhado dos manuais, que orientam o modelo com informações corretas e diretamente relacionadas ao funcionamento dos sistemas automotivos.

Além do desempenho dos modelos, foi realizada também uma avaliação do custo operacional associado a cada abordagem, considerando o volume total de *tokens* processados nos experimentos. Cada método foi aplicado ao mesmo conjunto de 492 amostras, composto por pares pergunta–resposta avaliados com ou sem a adição de contexto, garantindo comparabilidade direta entre os cenários. A Tabela 6 apresenta os valores médios de consumo de *tokens* por método, bem como o custo estimado com base nas tarifas públicas de uso das APIs do GPT-4o-mini e do Gemini 2.5-flash-lite ($0,15 por milhão de *tokens* para ambos os modelos).

Observa-se que o método *zero-shot* apresenta o menor

Modelo	top_k	Precisão	Recall	F1
GPT	1	69.41 ± 4.73	39.06 ± 1.93	49.91 ± 0.44
GPT	3	79.58 ± 10.69	41.19 ± 1.86	54.08 ± 1.37
GPT	5	67.83 ± 1.22	41.06 ± 2.01	51.12 ± 1.24
Gemini	1	57.43 ± 0.64	39.29 ± 0.27	46.66 ± 0.34
Gemini	3	60.86 ± 0.58	44.04 ± 0.96	51.10 ± 0.80
Gemini	5	55.87 ± 0.96	41.33 ± 0.96	47.51 ± 0.55

Tabela 4. Desempenho no cenário *Few-shot* para diferentes valores de top_k

Modelo	top_k	Precisão	Recall	F1
GPT	1	87.14 ± 5.44	59.62 ± 4.72	70.64 ± 1.98
GPT	3	89.95 ± 3.39	62.26 ± 3.49	73.51 ± 1.37
GPT	5	85.45 ± 1.88	64.36 ± 0.26	73.42 ± 0.76
Gemini	1	87.96 ± 5.19	47.16 ± 3.01	61.30 ± 1.59
Gemini	3	89.84 ± 3.33	51.09 ± 1.33	65.11 ± 1.04
Gemini	5	85.18 ± 1.42	50.54 ± 0.70	63.43 ± 0.40

Tabela 5. Desempenho no cenário RAG-Manuais para diferentes valores de top_k

custo entre todas as abordagens, uma vez que utiliza unicamente o par pergunta–resposta, sem qualquer adição de contexto. O consumo cresce gradualmente nas variantes de *few-shot*, devido ao aumento proporcional do tamanho do contexto de entrada conforme o valor de *top_k*. Nas abordagens baseadas em manuais técnicos, os custos são mais elevados, especialmente nos cenários com maiores valores de *top_k*, o que reflete diretamente o tamanho substancial dos trechos provenientes dos manuais.

Apesar do custo mais elevado, o *RAG-Manuais* apresentou o melhor desempenho entre todos os métodos, o que indica um potencial equilíbrio entre custo e benefício. Esse resultado reforça que, embora consumir mais *tokens* implique maior custo financeiro, a utilização de documentação técnica resulta em ganhos expressivos de desempenho que podem justificar a despesa adicional, especialmente em sistemas que priorizam precisão e confiabilidade na validação de respostas técnicas.

Esses resultados sugerem que, embora exista uma correlação direta entre a quantidade de contexto e o custo computacional, o ganho de desempenho obtido com o uso de documentação técnica formal pode compensar o aumento de gasto, dependendo do cenário de aplicação. Em ambientes de validação automática em larga escala, a escolha entre custo e precisão deve considerar o grau de confiabilidade esperado pelo sistema e o impacto potencial de erros na tomada de decisão. Diante desse conjunto de evidências, a seção seguinte discute de forma integrada os efeitos observados, interpretando as diferenças entre abordagens e modelos à luz das métricas e intervalos de confiança obtidos.

7 Discussão

A discussão a seguir analisa os resultados obtidos à luz das diferentes abordagens avaliadas, considerando o impacto da recuperação de contexto, o tipo de fonte utilizada e as diferenças observadas entre os modelos.

A análise comparativa entre os três métodos evidencia uma progressão consistente no desempenho dos modelos, o que também se reflete nas métricas consolidadas e nos intervalos de confiança. O cenário *zero-shot* apresenta os menores valores de F1 e a maior variabilidade, refletindo a limitação dos modelos ao operarem sem qualquer apoio contextual. Nesse caso, ambos os modelos exibem desempenho substancialmente inferior, com destaque para a maior incerteza observada no GPT, cujo intervalo de confiança é mais amplo.

A introdução de exemplos de perguntas semanticamente semelhantes, por meio do método *few-shot*, resulta em um avanço moderado em precisão e F1, indicando que a recuperação de informações desempenha papel relevante na tarefa de validação. Essa abordagem reduz ambiguidades ao expor o modelo a casos relacionados, mas permanece limitada pela própria natureza das discussões humanas presentes no fórum, que podem incluir respostas incompletas, imprecisas ou ambíguas. Essa característica explica a estabilidade intermediária observada nos intervalos de confiança e o desempenho inferior em comparação ao uso de documentação técnica.

O maior ganho é observado no método *RAG-Manuais*, no qual a recuperação de trechos provenientes de manuais automotivos fornece suporte substancialmente mais preciso e confiável. A presença de documentação estruturada e formalmente validada aumenta de maneira significativa o *recall* e reduz o desvio entre grupos, resultando em métricas superiores e intervalos de confiança mais estreitos. Esses achados reforçam a hipótese de que contextos técnicos normativos oferecem suporte mais robusto para a verificação automática, diferentemente do conteúdo heterogêneo recuperado de discussões entre usuários. Esse comportamento também está alinhado à literatura, que aponta benefícios semelhantes ao integrar LLMs com documentação técnica especializada, como discutido em Medeiros et al. [2023].

Além disso, observa-se que o GPT supera o Gemini em todos os cenários avaliados. A diferença é especialmente evidente no método *RAG-Manuais*, em que o GPT atinge valores de F1 acima de 70%, enquanto o Gemini permanece próximo de 63%. A baixa sobreposição entre os intervalos indica que essa diferença não é atribuída apenas à variação amostral, mas representa uma vantagem consistente do GPT na tarefa proposta.

Por outro lado, as diferenças observadas entre os métodos

Método	Tokens Totais	Tokens Médios	Custo (US$)
Zero-shot	95 612	194,33	0.0143
Few-shot (k=1)	152 463	309,88	0.0229
Few-shot (k=3)	270 945	550,70	0.0406
Few-shot (k=5)	392 470	797,70	0.0589
RAG-Manuais (k=1)	348 198	707,72	0.0522
RAG-Manuais (k=3)	850 538	1 728,74	0.1276
RAG-Manuais (k=5)	1 352 895	2 749,79	0.2029

Tabela 6. Consumo total de *tokens* e custo estimado por abordagem (tarifa de $0,15 por milhão de *tokens*).

não se limitam às métricas finais, mas também à natureza das respostas produzidas. No cenário *zero-shot*, parte do desempenho inferior parece decorrer de vieses intrínsecos dos modelos, que tendem a gerar respostas linguisticamente plausíveis mesmo quando incorretas, sobretudo na ausência de um contexto técnico adequado [Brown et al., 2020]. Esse fenômeno ajuda a explicar a forte variabilidade observada nos resultados.

Embora os achados sejam consistentes, algumas limitações devem ser consideradas. O número reduzido de grupos utilizados para estimar os intervalos de confiança limita a robustez estatística das estimativas, ainda que permita uma comparação inicial confiável entre abordagens. Adicionalmente, o método *few-shot* depende da qualidade das respostas humanas presentes no fórum, o que introduz ruído semântico e pode comprometer o contexto recuperado.

Outro ponto diz respeito aos modelos avaliados, que representam versões específicas das arquiteturas GPT e Gemini; versões mais recentes podem exibir comportamentos distintos. Por fim, o escopo deste estudo está restrito ao domínio automotivo e ao conjunto de manuais disponíveis, o que pode limitar a generalização dos resultados para outros setores técnicos. Ainda assim, os achados fornecem evidências claras sobre o papel decisivo do suporte contextual estruturado na melhoria do desempenho dos modelos de linguagem.

8 Conclusão

Este estudo avaliou a capacidade de modelos de linguagem de validar respostas técnicas automotivas por meio de três abordagens distintas: *zero-shot*, *few-shot* e RAG-Manuais. Os resultados demonstraram que a recuperação de contexto exerce papel decisivo no desempenho, com destaque para a abordagem baseada em manuais, que apresentou os maiores valores de F1-score e intervalos de confiança mais estreitos. Esses achados reforçam que, em domínios altamente técnicos, o acesso a documentação especializada é mais determinante do que o conhecimento interno do modelo.

A comparação entre modelos revelou diferenças consistentes, porém, ambos se beneficiaram das estratégias de RAG, indicando que a limitação principal não está no modelo em si, mas na disponibilidade e qualidade do contexto fornecido. Assim, o trabalho evidencia o potencial dos LLMs como ferramentas de apoio à verificação técnica, desde que operem em integração com fontes externas confiáveis.

Em conjunto, os resultados confirmam a viabilidade da metodologia proposta e apontam para aplicações reais em suporte técnico, triagem de respostas e sistemas de auxílio ao diagnóstico. A investigação abre caminho para aprimoramentos futuros, incluindo ampliação das bases documentais, novos mecanismos de recuperação e avaliação em cenários mais complexos.

Embora os resultados demonstrem a viabilidade da abordagem proposta, diversos aprimoramentos podem ser explorados em estudos posteriores. Entre as possibilidades, destaca-se a avaliação de modelos mais recentes e de maior capacidade, bem como a investigação de métodos alternativos de recuperação, incluindo técnicas baseadas em similaridade multidimensional ou re-ranking supervisionado, estratégia que tem demonstrado ganhos substanciais na qualidade da recuperação em diferentes domínios [Nogueira and Cho, 2019]. Outra direção promissora envolve a experimentação com diferentes estratégias de segmentação e de representação dos manuais, avaliando seu impacto direto na precisão dos modelos. Adicionalmente, a criação de um conjunto anotado maior e mais diversificado pode reduzir a variabilidade entre grupos e permitir análises estatísticas mais robustas. Por fim, uma extensão natural deste trabalho seria o desenvolvimento de um sistema aplicado, capaz de auxiliar profissionais da área automotiva na validação de respostas em tempo real.

Referências

Tom B Brown, Benjamin Mann, and Nick et al. Ryder. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Ilias Chalkidis, Manos Fergadiotis, Panagiotis Mangonas, and Ion Androutsopoulos. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of ACL*, 2022.

Google DeepMind. Gemini model documentation. `https://deepmind.google/`, 2023.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. A deep look into neural ranking models for information retrieval. In *Information Retrieval Journal*, volume 23, pages 273–345, 2020.

Kelvin Guu, Kenton Lee, Zora Tung Chang, and Tom Kwiatkowski. Retrieval-augmented language model pretraining. In *Proceedings of ICML*, 2020.

Ziwei Ji et al. A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. In *IEEE Transactions on Big Data*, pages 1–12, 2019. .

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Wei-Cheng Ko, Pragyan Radhakrishnan, and Mona Diab. Large language models for context-aware evaluation of technical question answering. In *Proceedings of EMNLP*, 2023.

Patrick Lewis, Ethan Perez, and Aleksandra et al. Piktus. Retrieval-augmented generation for knowledge-intensive nlp. In *Advances in Neural Information Processing Systems*, 2020.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Thaís Medeiros, Morsinaldo Medeiros, Mariana Azevedo, Marianne Silva, Ivanovitch Silva, and Daniel G. Costa. Analysis of language-model-powered chatbots for query resolution in pdf-based automotive manuals. *Vehicles*, 5 (4):1384–1399, 2023. ISSN 2624-8921. . URL `https://www.mdpi.com/2624-8921/5/4/76`.

Niklas Muennighoff et al. Mteb: Massive text embedding benchmark. In *NeurIPS*, 2022.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. In *arXiv preprint arXiv:1901.04085*, 2019.

OpenAI. Openai embeddings documentation. `https://platform.openai.com/docs`, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

Ankit Singh, Rishabh Gupta, and Varun Kumar. Medical-rag: A retrieval-augmented generation framework for medical question answering. In *Proceedings of the EMNLP Workshop on Healthcare NLP*, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: A large-scale dataset for fact extraction and verification. In *Proceedings of NAACL-HLT*, pages 809–819, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jie Zhou, Tao Zhang, and Jidong Wang. A comprehensive evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.