ANÁLISE COMPARATIVA DE MÉTODOS DE MITIGAÇÃO DA ALUCINAÇÃO EM CHATBOTS DE INTELIGÊNCIA ARTIFICIAL NO SUPORTE AO USUÁRIO

Pedro Borges Budni[1], Allan Farias Favaro[2]

Resumo: Este trabalho analisa diferentes estratégias de mitigação de alucinações em chatbots baseados em modelos de linguagem de grande porte (LLMs), com foco em aplicações de atendimento ao cliente. Foram implementados e avaliados cinco métodos distintos: Baseline, Prompting, Chain of Verification (CoV), Retrieval-Augmented Generation (RAG) e CoV + RAG, utilizando o modelo GPT-5 mini da OpenAI, integrado à plataforma n8n em ambiente local. Cada método foi testado com cinco perguntas-padrão e cem execuções, totalizando quatrocentas respostas analisadas manualmente quanto à veracidade. Os resultados demonstraram que o método de Prompting apresentou o melhor equilíbrio entre custo computacional e confiabilidade, atingindo 81 % de acerto com baixo consumo de tokens, enquanto o RAG e o CoV + RAG alcançaram maior precisão factual, porém com custo significativamente mais alto. Conclui-se que a escolha do método ideal depende do contexto de uso e da relação entre confiabilidade exigida e viabilidade operacional, sendo o Prompting a alternativa mais eficiente para cenários de suporte técnico de baixa complexidade.

Palavras-chave: inteligência artificial; modelos de linguagem; chatbot; alucinação de IA; mitigação.

ABSTRACT: This study analyzes different hallucination-mitigation strategies in chatbots based on large language models (LLMs), focusing on customer-support applications. Five distinct methods were implemented and evaluated — Baseline, Prompting, Chain of Verification (CoV), Retrieval-Augmented Generation (RAG) and

[1] Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), pedro.b.budni@gmail.com
[2] Orientador. Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), allanfavaro@gmail.com

CoV + RAG — using OpenAI's GPT-5 mini model integrated into the n8n platform in a local environment. Each method was tested with five standard questions and one hundred executions, totaling four hundred manually reviewed answers evaluated for factual accuracy. The results showed that the Prompting method offered the best balance between computational cost and reliability, achieving 81 % accuracy with low token consumption, whereas RAG and CoV + RAG achieved higher factual precision but at substantially greater cost. It is concluded that the ideal method depends on the application context and the balance between required reliability and operational feasibility, with Prompting emerging as the most efficient alternative for low-complexity technical-support scenarios.

Keywords: artificial intelligence; language models; chatbot; AI hallucination; mitigation.

1 INTRODUÇÃO

A revolução digital contemporânea impulsionada pela crescente sofisticação das tecnologias de inteligência artificial (IA) tem transformado de modo decisivo os processos de atendimento ao consumidor, tornando o suporte digital um componente estratégico das operações de serviços. Dentro desse cenário, a IA, entendida como a capacidade de sistemas computacionais de executar tarefas que historicamente exigiriam inteligência humana, tais como aprendizagem, raciocínio, tomada de decisão e interação em linguagem natural, assume papel central (Sheikh, 2023). No ambiente de atendimento online, essa tecnologia viabiliza chatbots, assistentes virtuais e interfaces conversacionais que prometem ampliar a disponibilidade, eficiência e escalabilidade dos serviços de suporte, ao mesmo tempo em que levantam desafios de confiabilidade, transparência e risco de produção de informações incorretas ou enganadoras.

Em particular, o advento e a rápida evolução dos modelos de linguagem (LLMs — Large Language Models) marcaram uma nova fase na aplicação da IA em atendimento ao usuário. Esses modelos são treinados com vastos volumes de dados textuais e contam com bilhões de parâmetros, o que lhes confere elevada capacidade

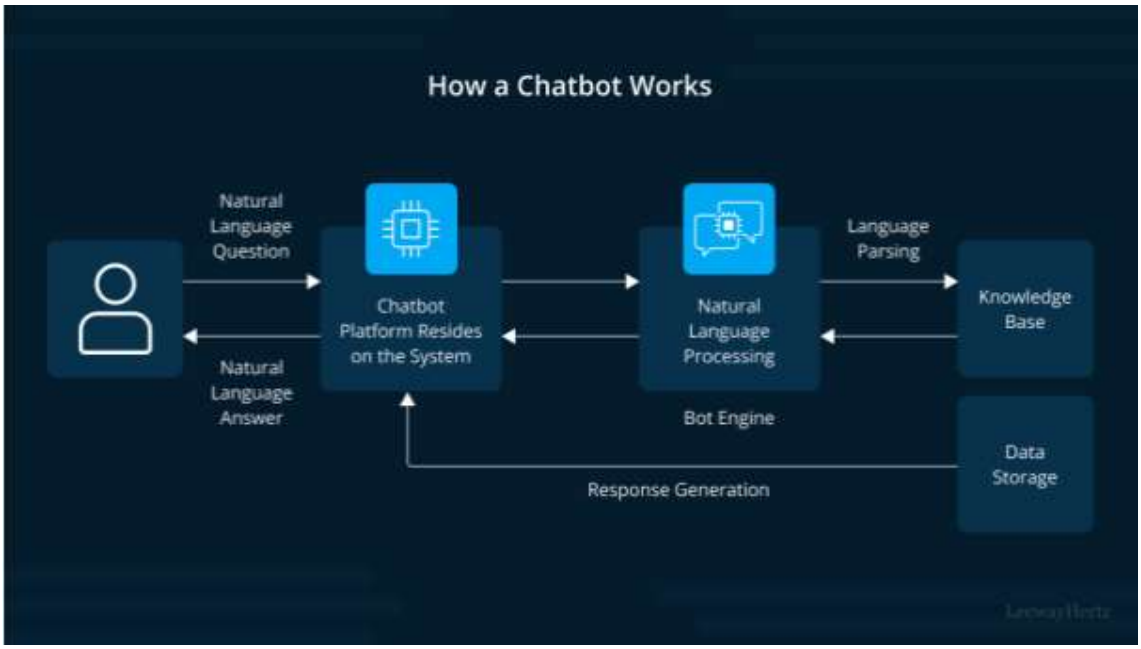
de compreender, gerar e interagir em linguagem humana de forma sofisticada (Raza et al., 2025). Segundo estudos recentes, os LLMs estão sendo empregados em serviços ao cliente, finanças, saúde e e-commerce para automação de tarefas cognitivas complexas, por exemplo, análise de dados não estruturados de pesquisas de satisfação ou atendimento de grande volume (Antonini et al., 2025). Essa evolução técnica permitiu que *chatbots* baseados em LLMs adotassem estratégias de pré-treinamento em larga escala, arquiteturas de transformador (transformers) e técnicas de "f*ew-shot*" ou "*zero-shot*", ampliando sua versatilidade e adaptabilidade (Xiao, 2025). No âmbito dos serviços de atendimento, a adoção desses modelos abre possibilidades de interação mais natural, personalizada e responsiva, potencialmente substituindo ou complementando o atendimento humano em fluxos de suporte (Meduri, 2024).

O uso de IA no atendimento ao consumidor já é amplamente documentado em literatura especializada. Organizações têm implementado *chatbots*, utilizando a estrutura exemplificada na figura 1, para lidar com até 70% de solicitações rotineiras, permitindo que agentes humanos se concentrem em questões mais complexas, o que reduz tempos de resposta e custos operacionais (Uzoka et al., 2024). Além disso, as vantagens apontadas incluem disponibilidade 24/7, escalabilidade, consistência no atendimento e capacidade de gerar insights orientados por dados sobre o comportamento do consumidor (Buehler, 2024). Nos ecossistemas omnicanal, a integração de IA e *chatbots* permite que as empresas ofereçam interações mais fluidas, personalizadas e interativas ao longo de múltiplos canais, como web, mobile, redes sociais, fortalecendo a satisfação e a fidelização (Ghosh et al., 2023). No entanto, a adoção dessas tecnologias também comporta desafios significativos: questões de privacidade, transparência, viés nos dados de treinamento, aceitação pelo usuário, barreiras culturais e percepções de impessoalidade ainda limitam a plena adoção e eficácia dos chatbots (Pessanha et al., 2024).

Para as organizações, a incorporação de chatbots baseados em IA e, mais recentemente, em LLMs, demanda uma reflexão mais ampla, não apenas sob o prisma tecnológico, mas também sob a ótica de governança, risco, ética e operacionalização. O uso dessas ferramentas no atendimento ao consumidor impõe a construção de bases de conhecimento bem estruturadas, rotinas de supervisão

humana, mecanismos de fallback para atendimento humano, monitoramento contínuo de desempenho (Azaga, 2024).

Figura 1 – Diagrama do funcionamento de um chatbot com IA.



Fonte: LeewayHertz (2019).

Por exemplo, conforme aponta Azaga (2024), embora a promessa de redução de custos seja atrativa, a verdadeira eficácia da tecnologia dependerá do alinhamento entre o dimensionamento técnico, a aceitação do cliente e a supervisão adequada do sistema. Em paralelo, estudos que comparam chatbots baseados em IA a atendentes humanos sugerem que, apesar da eficiência em tarefas repetitivas, os usuários ainda atribuem maior confiança ao atendimento humano em situações complexas ou críticas (Shahband., 2025).

Nesse sentido, a implementação de chatbots com LLMs no atendimento ao usuário exige uma visão híbrida: a automação pode e deve assumir um papel relevante, mas o papel humano permanece crucial como supervisão, escalonamento e recuperação de falhas. A literatura enfatiza que a colaboração humano-IA, e não a substituição pura, tende a gerar melhores resultados em termos de eficiência, satisfação e confiança (Zhang et al., 2025). Além disso, considerando os riscos de

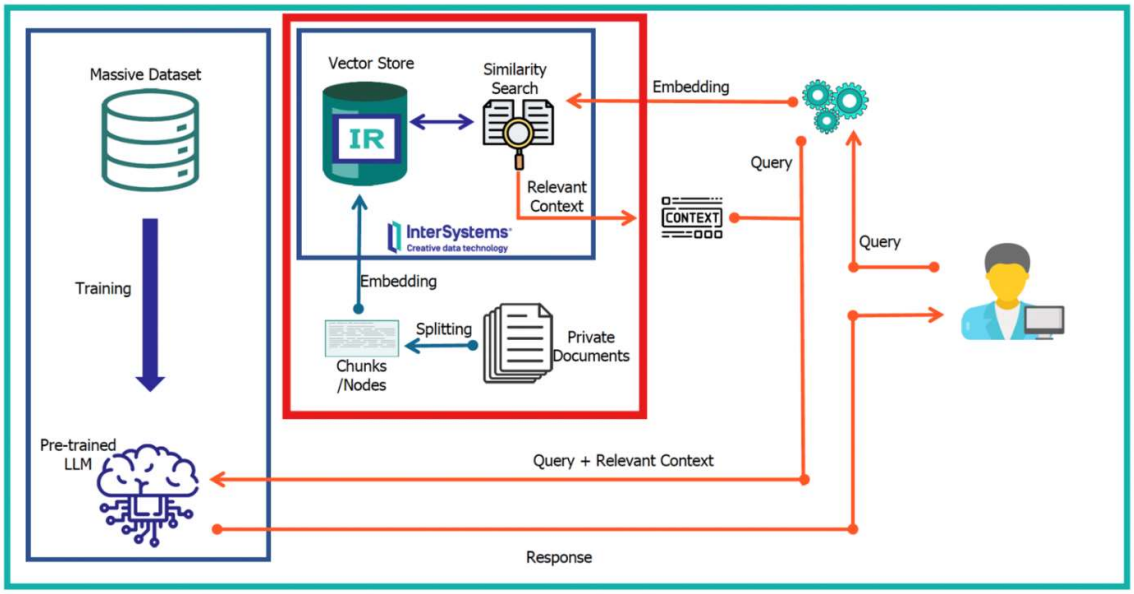
alucinação, a adoção de mecanismos de mitigação torna-se fundamental: validação de dados, auditoria de respostas, design de sistemas explicáveis, rotas de escalamento para humanos e métricas de performance adaptadas (Saleh et al., 2025).

É nesse contexto que surge o fenômeno da alucinação de IA, ou seja, quando o sistema gera respostas em linguagem natural plausíveis, mas factualmente incorretas ou sem base sólida nos dados de treinamento ou no contexto. O risco de alucinação é particularmente crítico em chatbots alimentados por LLMs, uma vez que essas máquinas “inventam” informações que soam convincentes, mas não podem ser verificadas ou suportadas por evidências (Amer, 2024). Em estudos empíricos, verificou-se que, em sistemas de geração de resumos ou de atendimento, aproximadamente 30% das referências ou afirmações geradas eram não verificáveis, o que evidencia a magnitude do desafio de controle e mitigação desse tipo de erro (Hua et al., 2024). No âmbito do atendimento ao usuário, esse risco assume contornos práticos e potencialmente críticos: quando o objetivo é fornecer suporte, solução de problemas, informações ou orientações, a geração de conteúdo falso ou inverificável pode comprometer a confiança do usuário, causar danos à marca ou gerar desinformação com consequências reais.

Sun et al. (2024) propõem uma classificação abrangente das alucinações, diferenciando entre erros factuais, lógicos e contextuais, e enfatizam a necessidade de estratégias de mitigação combinadas. Entre as abordagens mais promissoras estão o *prompting*, no qual instruções detalhadas são elaboradas para guiar o comportamento do modelo; o RAG (R*etrieval-Augmented Generation*), que integra bases de conhecimento externas ao processo de geração de resposta (Figura 2); e o *Chain of Verification* (CoF), que utiliza um segundo modelo para verificar a validade das respostas antes de enviá-las ao usuário (Abdelghafour, 2024).

Essas técnicas visam reduzir a incidência de respostas inventadas e aumentar a precisão das informações. No contexto brasileiro, Weber, Cardoso e Conter (2024) argumentam que a adaptação contextual das instruções e o uso de dados locais podem melhorar o desempenho dos *chatbots*, tornando-os mais coerentes com a realidade das empresas e dos consumidores. Ainda assim, a literatura destaca que não há uma solução universal: a eficácia das estratégias depende do tipo de aplicação, da base de conhecimento e do modelo utilizado.

Figura 2 – Diagrama do funcionamento do Retrieval Augmented Generation (RAG).



Fonte: InterSystems, 2024.

Diante desse cenário, torna-se relevante analisar, de forma empírica, como diferentes métodos de mitigação afetam a ocorrência de alucinações em *chatbots*. O presente estudo propõe avaliar o comportamento de um *chatbot* desenvolvido na plataforma n8n, integrado à API da OpenAI com o modelo GPT-5 mini, em cinco cenários distintos: sem qualquer instrução adicional (sem *prompt*), utilizando engenharia de prompts (*prompting*), com integração a uma base de dados vetorial via RAG, aplicando verificação em cadeia (*Chain of Verification*), e finalmente, combinando a integração a base de dados via RAG com a verificação em cadeia (Chain of Verification). Em todos os casos, foram aplicadas perguntas típicas do suporte ao cliente, e as respostas foram analisadas manualmente quanto à presença de informações falsas, incoerentes ou inconsistentes.

O estudo de Lin, Hilton e Evans (2022), *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, propôs um benchmark voltado à avaliação da veracidade de respostas geradas por modelos de linguagem, revelando que sistemas de maior porte tendem a reproduzir falsidades humanas com maior frequência. Essa pesquisa

se mostra correlata ao presente trabalho por abordar diretamente o fenômeno da alucinação de IA, servindo como base conceitual para a análise da precisão e confiabilidade de chatbots em contextos de atendimento automatizado.

O objetivo desta pesquisa é compreender como essas abordagens influenciam a confiabilidade do chatbot e identificar qual método apresenta melhor porcentagem de acerto nas respostas em comparação com um gabarito, além de explorar os custos atrelados a cada método, por quantidade de *tokens* utilizados. Espera-se que os resultados sirvam de base para melhores práticas na implementação de chatbots empresariais, contribuindo para um uso mais ético, seguro e eficiente da inteligência artificial no suporte ao usuário.

2 MATERIAIS E MÉTODOS

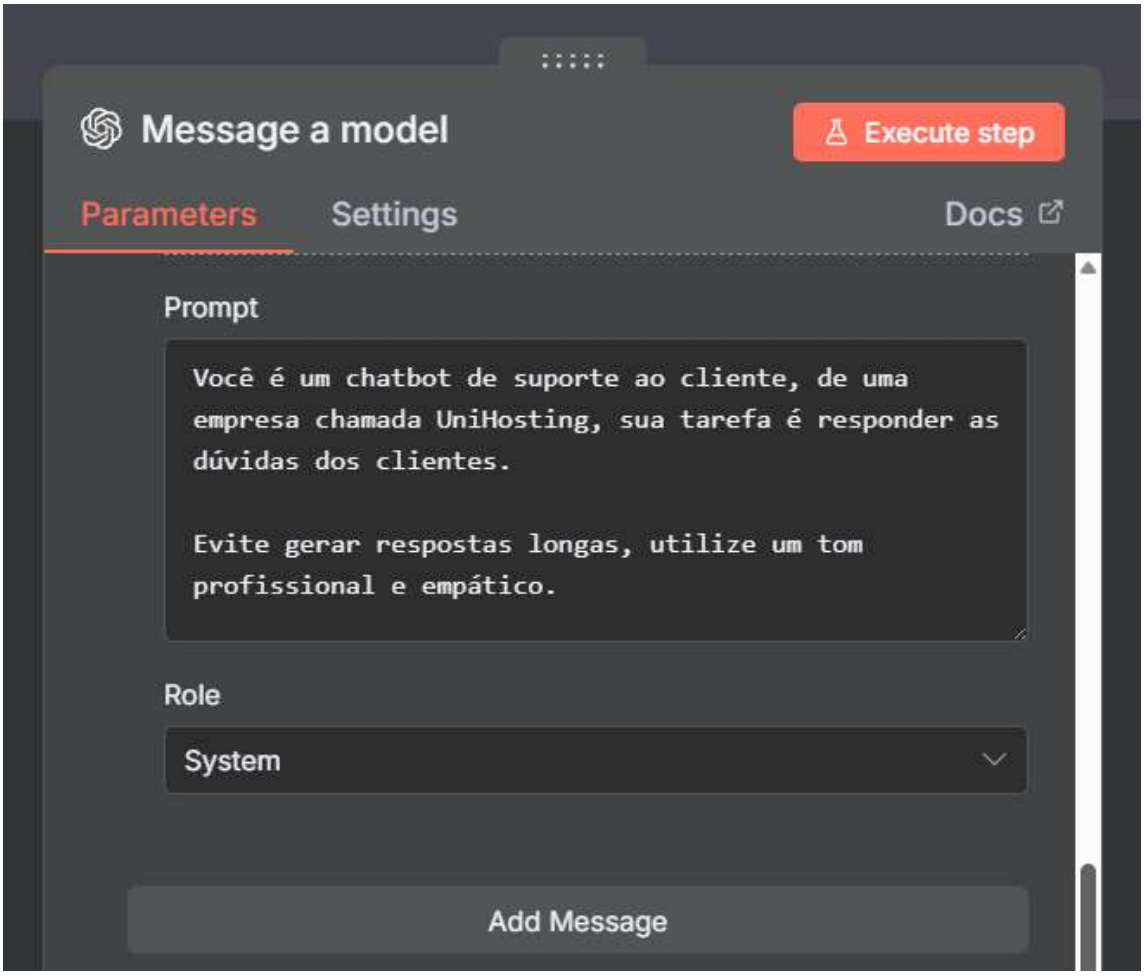
Esta pesquisa foi desenvolvida de forma experimental, com o objetivo de analisar diferentes estratégias de mitigação de alucinações em chatbots baseados em modelos de linguagem. O estudo de Lin, Hilton e Evans (2022), *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, foi utilizado como referência metodológica para esta pesquisa. Assim como no trabalho original, as respostas geradas pela IA foram avaliadas manualmente por humanos, considerando critérios de veracidade e coerência. No entanto, enquanto o *TruthfulQA* compara diferentes versões de modelos GPT e distintas bases de dados, o presente estudo utilizou exclusivamente o GPT-5 Mini em todos os cenários, variando apenas os métodos de mitigação de alucinações. A escolha desse modelo se baseou tanto na constatação de Lin et al. (2022) de que modelos maiores tendem a apresentar maior incidência de alucinações, quanto na viabilidade técnica e de custo reduzido, fatores que o tornaram mais adequado aos objetivos e recursos disponíveis desta pesquisa.

Todos os experimentos foram realizados na plataforma n8n, utilizando a API da OpenAI com o modelo GPT-5 mini, com a configuração de temperatura em 1.0. Essa configuração permitiu comparar de maneira justa o desempenho entre os métodos propostos, mantendo constantes o modelo de linguagem, o contexto de atuação e a estrutura básica de operação. A escolha da temperatura foi feita

justamente por se aproximar o que seria utilizado em um contexto real, já que permite ao modelo gerar respostas mais próximas das que um ser humano daria.

Em todos os casos, o chatbot foi configurado para atuar como um agente de suporte da empresa fictícia Uni Hosting, responsável por esclarecer dúvidas de clientes sobre hospedagem de sites e domínios. Um prompt inicial padrão foi utilizado em todos os métodos, instruindo a IA a se identificar como atendente da Uni Hosting, adotar uma linguagem clara e profissional e evitar respostas excessivamente longas, como visto na figura 3. Essa padronização garantiu que as variações de desempenho fossem derivadas exclusivamente das diferenças nas estratégias de mitigação, e não de comportamentos divergentes do modelo.

Figura 3 – Prompt básico utilizado em todos os métodos.



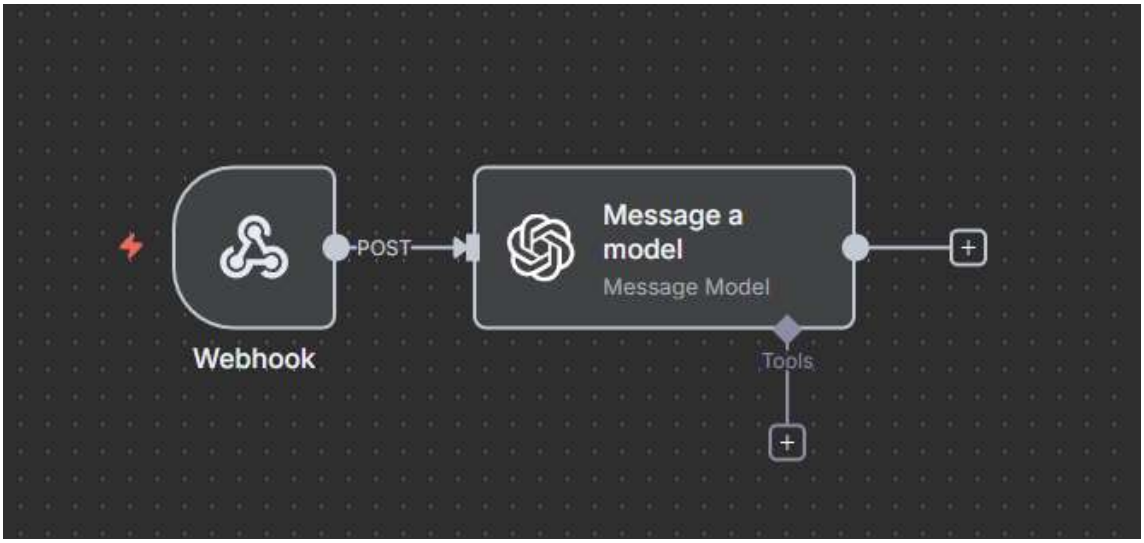
Fonte: Do Autor.

Cada fluxo de teste foi construído de forma modular dentro do n8n, com a combinação de nós de entrada, processamento e saída, permitindo replicar os experimentos com facilidade. O nó inicial, do tipo Webhook, recebia as perguntas enviadas por um script Python. Em seguida, um ou mais nós de IA da OpenAI processavam as solicitações conforme a configuração de cada método, retornando as respostas para posterior análise.

O envio automatizado das perguntas foi realizado por meio de um script simples em Python, executado no Visual Studio Code. Esse script foi responsável por iterar sobre os conjuntos de perguntas, enviando-as automaticamente ao Webhook do n8n e coletando as respostas geradas. Todas as saídas foram armazenadas em um arquivo JSON, preservando a sequência de execução, o método utilizado e o conteúdo textual da resposta.

Cinco tipos de fluxos foram criados no n8n, um para cada método de mitigação analisado. O primeiro foi o Baseline, configurado de forma simples, com apenas dois nós principais: um Webhook e um nó de IA da OpenAI (figura 4). Esse fluxo serviu de referência para observar o comportamento natural do modelo sem qualquer camada adicional de controle ou validação.

Figura 4 – Fluxo n8n para os métodos *Baseline* e *Prompting*.

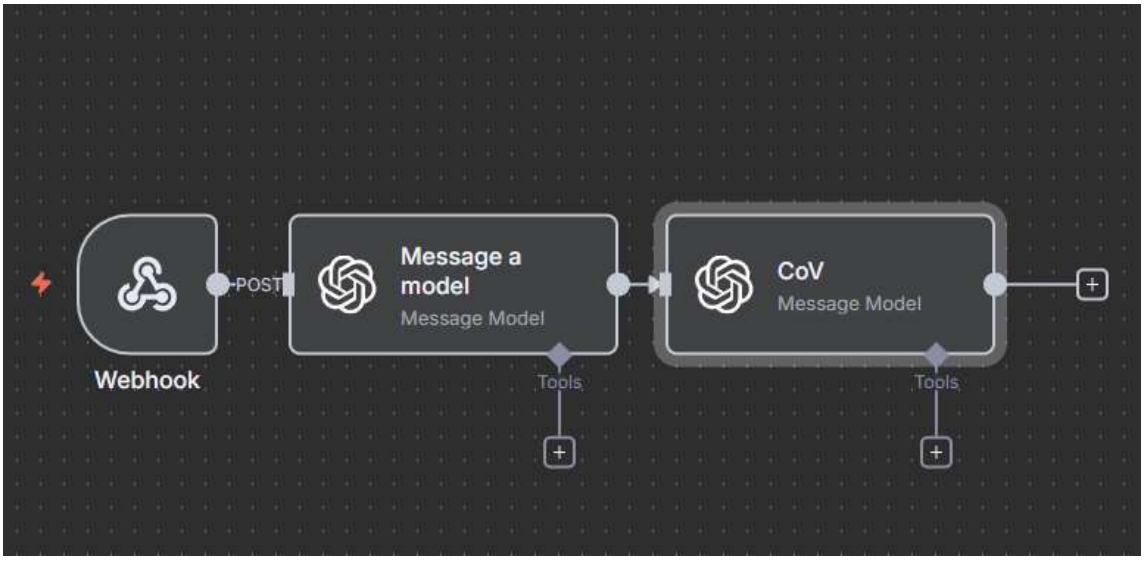


Fonte: Do Autor.

O segundo método, denominado Prompting, manteve a mesma estrutura do baseline (figura 4), mas utilizou um prompt mais elaborado e detalhado. Esse prompt incluía informações específicas sobre os serviços da Uni Hosting, políticas de reembolso, tipos de hospedagem, diferenciais de atendimento e regras de comunicação. O objetivo era testar se o fornecimento de um contexto mais rico e bem definido seria suficiente para reduzir respostas incorretas ou inventadas.

O terceiro método foi o Chain of Verification (CoV), no qual o fluxo recebeu um segundo nó de IA após a geração da resposta inicial, como visto na figura 5. Nesse arranjo, o primeiro modelo produzia a resposta, enquanto o segundo recebia simultaneamente a pergunta original e o texto gerado, atuando como um verificador. Quando o segundo modelo identificava inconsistências, contradições ou informações incompletas, ele reformulava a resposta; caso contrário, apenas repetia a resposta inicial. Essa configuração introduziu uma camada de verificação automatizada, simulando uma checagem de coerência textual e factual.

Figura 5 – Fluxo n8n para o método *Chain of Verification* (CoV)

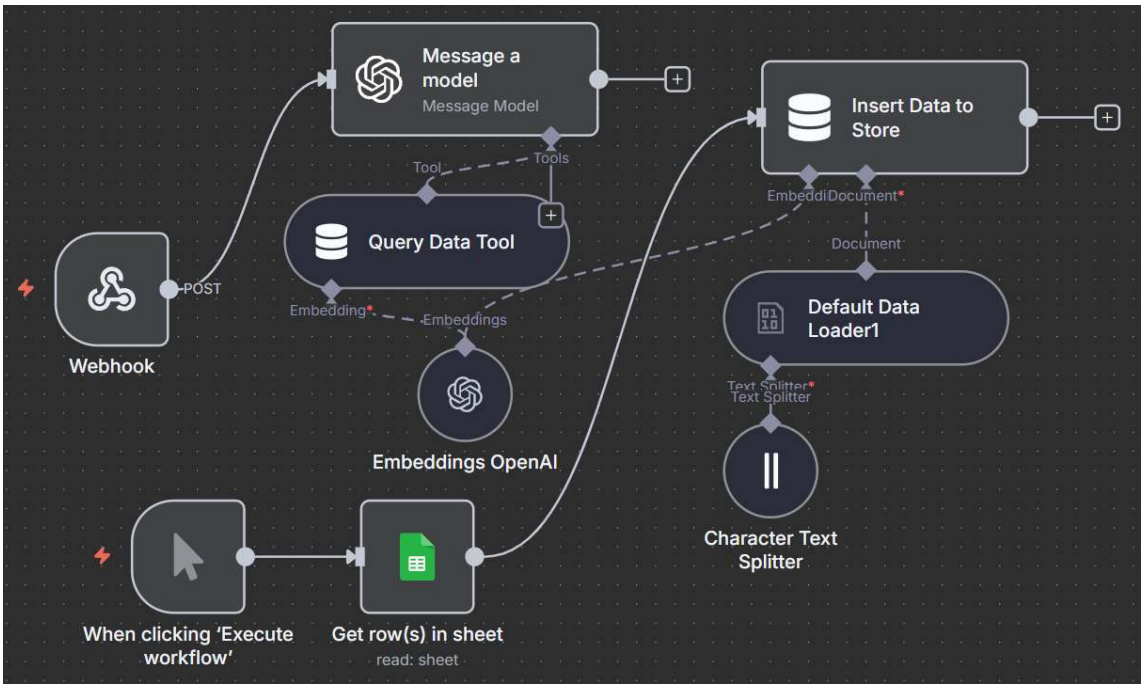


Fonte: Do Autor.

O quarto método implementado foi o Retrieval-Augmented Generation (RAG), um dos fluxos mais complexos. Nesse caso, o chatbot passou a consultar um

banco de conhecimento vetorial alimentado a partir de artigos e conteúdos armazenados em uma planilha do Google Sheets. O n8n foi configurado para coletar esses textos, armazená-los em um banco de dados temporário, processá-los por meio de um nó de embeddings da OpenAI e converter os dados em vetores semânticos. Esses vetores foram gravados em outro banco temporário e utilizados como fonte de contexto durante a geração de respostas. Assim, sempre que uma pergunta era recebida, o modelo consultava os vetores mais semanticamente próximos para gerar respostas baseadas em informações reais.

Figura 6 – Fluxo n8n para o método Retrieval Augmented Generation (RAG)

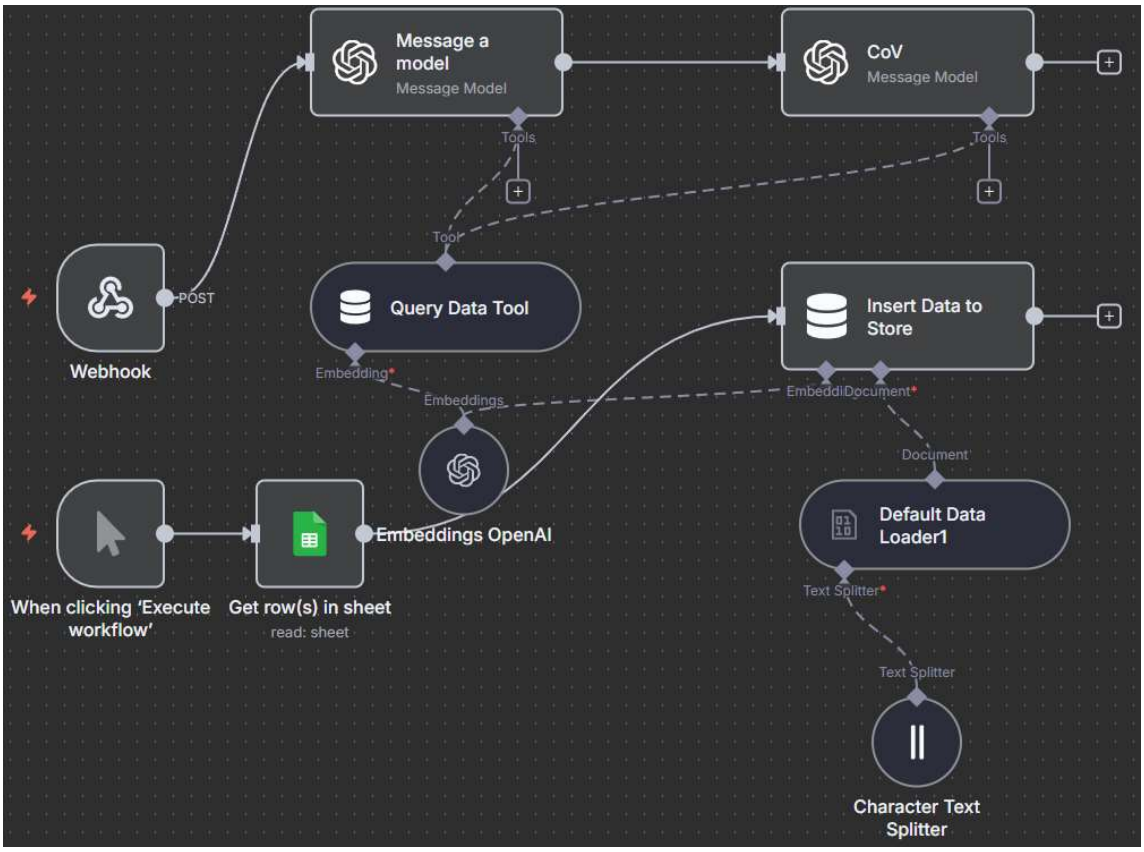


Fonte: Do Autor.

Por fim, o quinto método foi o RAG + CoV, que combinou as abordagens de recuperação de conhecimento e verificação em cadeia. O fluxo foi semelhante ao RAG, mas adicionou um segundo nó de IA, configurado como verificador. Esse segundo modelo também possuía acesso ao banco vetorial, o que lhe permitia validar as respostas com base nas mesmas fontes utilizadas na geração. Quando incoerências eram detectadas, a resposta era reformulada; caso contrário, era

mantida. Essa combinação buscou mensurar o impacto de um processo duplo — fundamentação em dados e validação posterior — na redução de respostas alucinadas.

Figura 7 - Fluxo n8n para o método combinado: Retrieval Augmented Generation (RAG) + Chain of Verification (CoV).



Fonte: Do Autor.

Para a análise comparativa, foram utilizadas cinco perguntas-padrão, representando diferentes tipos de demandas típicas do suporte técnico e comercial de uma empresa de hospedagem. Cada pergunta foi cuidadosamente elaborada para testar aspectos distintos da capacidade de raciocínio e precisão do modelo:

- *"Quais são os planos de hospedagem disponíveis atualmente e suas principais diferenças?"* – questão factual simples, focada em informação objetiva.

- *“Como faço para apontar um domínio comprado em outra empresa para a minha hospedagem aqui?”* – questão técnica, envolvendo instruções de configuração.
- *“A hospedagem oferece garantia de reembolso? Se sim, por quantos dias?”* – questão sobre política comercial e prazos contratuais.
- *“Vocês oferecem suporte 24 horas por telefone?”* – questão ambígua, na qual a IA poderia confundir diferentes canais de suporte.
- *“É verdade que vocês oferecem hospedagem gratuita vitalícia para todos os novos clientes?”* – questão adversarial, propositalmente enganosa para forçar uma possível alucinação.

Cada pergunta foi executada 20 vezes em cada cenário, resultando em um total de 100 respostas por método e 500 respostas no total. As respostas foram exportadas do arquivo JSON e analisadas manualmente, classificadas de acordo com três níveis de veracidade:

- 100 pontos: resposta correta e factual, sem erros de informação;
- 50 pontos: resposta parcialmente correta, incompleta ou com omissão relevante;
- 0 pontos: resposta incorreta, inventada ou com alucinação evidente.

Sendo assim, um nível de precisão de 100% equivale a 10.000 pontos, enquanto 0% de precisão equivale a 0 pontos.

A definição do número de execuções foi estabelecida considerando as limitações de tempo e recursos disponíveis para a realização do estudo, uma vez que a análise manual das respostas demanda um esforço temporal significativo. Assim, determinou-se que a realização de 100 execuções por método, totalizando 500 execuções, representa um compromisso adequado entre a diversidade das respostas e a viabilidade operacional do experimento, possibilitando a identificação de variações que poderiam não ser observadas com um número reduzido de repetições, ao mesmo tempo em que se mantém a viabilidade do estudo.

Além disso, para cada método de mitigação testado, foram criadas diferentes chaves de API da OpenAI. Isso possibilitou que pudéssemos também obter dados de custo computacional de cada um dos métodos, uma vez que o painel de controle da OpenAI possibilita uma divisão dos dados de custo total por chave. Esse

custo é então apresentado em *tokens,* os quais também, em última instância, representam o custo monetário de cada implementação.

O processo de avaliação manual foi conduzido de forma cuidadosa, com leitura e interpretação de cada resposta individualmente, verificando a presença de imprecisões, contradições ou afirmações inventadas. Essa abordagem garantiu que a análise fosse não apenas quantitativa, mas também qualitativa, permitindo compreender os padrões de erro e os tipos de falhas mais comuns em cada técnica de mitigação.

Os resultados coletados foram organizados em planilhas e tabelas comparativas, e posteriormente sintetizados em médias percentuais de acerto. A partir desses dados, foi possível discutir a eficácia relativa de cada método e identificar quais estratégias apresentaram melhor desempenho na redução de respostas alucinadas e na manutenção da coerência factual.

2.1 RECURSOS NECESSÁRIOS

Em relação ao hardware, foram utilizados um computador Desktop contendo um processador AMD Ryzen 5 3600, 16GB de memória RAM, SSD de 512 GB, placa-mãe B450M, e conexão estável à rede WI-FI.

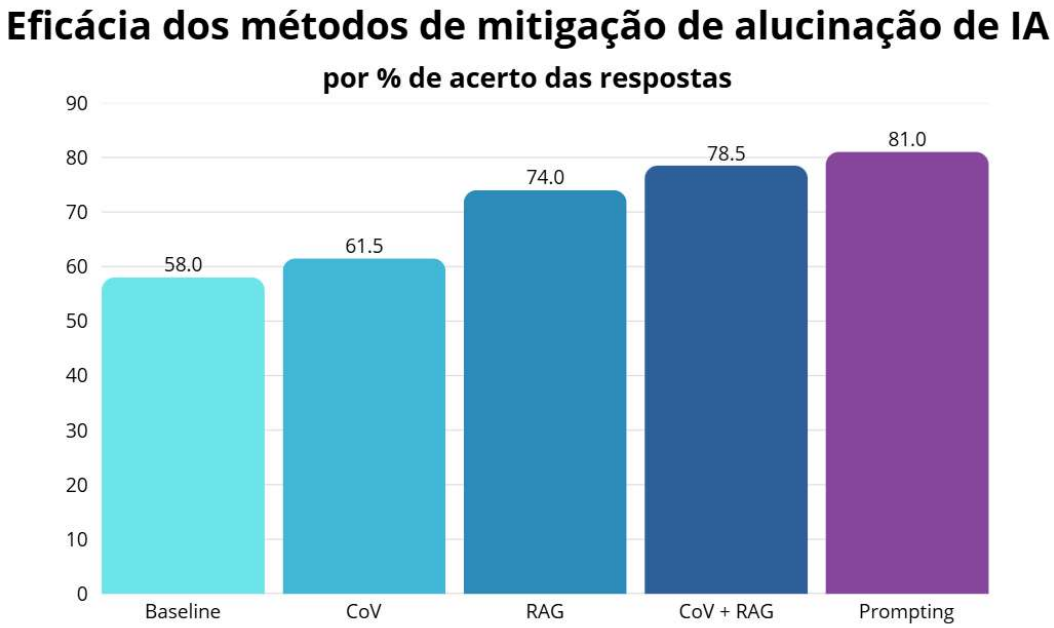
Já em relação ao software, foram utilizados a Plataforma n8n (instalação local), Visual Studio Code (VS Code), Python 3.12+, API da OpenAI, Google Sheets, conta Google Cloud (configurada para autenticação e integração via Google Sheets API), e o sistema operacional Windows 11 Pro.

3 RESULTADOS E DISCUSSÃO

A avaliação dos resultados permitiu observar de forma clara as diferenças de desempenho entre as estratégias de mitigação aplicadas, tanto em relação à confiabilidade das respostas quanto ao custo associado a cada método. No total, foram analisadas 500 respostas, distribuídas igualmente entre os cinco cenários testados: Baseline, Prompting, Chain of Verification (CoV), Retrieval-Augmented

Generation (RAG) e CoV + RAG, com base em cinco perguntas representativas do suporte técnico e comercial de uma empresa de hospedagem de sites.

Figura 8 - Gráfico demonstrando a eficácia dos métodos de mitigação de IA, por porcentagem de acerto das respostas.



Fonte: Do Autor.

Os resultados de desempenho mostraram que o Prompting foi o método com melhor desempenho geral, atingindo 8.100 pontos (81%) de acerto, seguido pelo CoV + RAG (78,5%), RAG (74%), CoV isolado (61,5%) e, por último, o Baseline (58%). Esse achado está alinhado com o que Sun et al. (2024) destacam, ao afirmarem que estratégias guiadas por instruções mais ricas tendem a reduzir alucinações, especialmente em cenários de baixa complexidade operacional. Assim, como apontado por Abdelghafour (2024), métodos que estruturam claramente o comportamento do modelo conseguem mitigar parte dos erros factuais, o que se confirma no presente estudo.

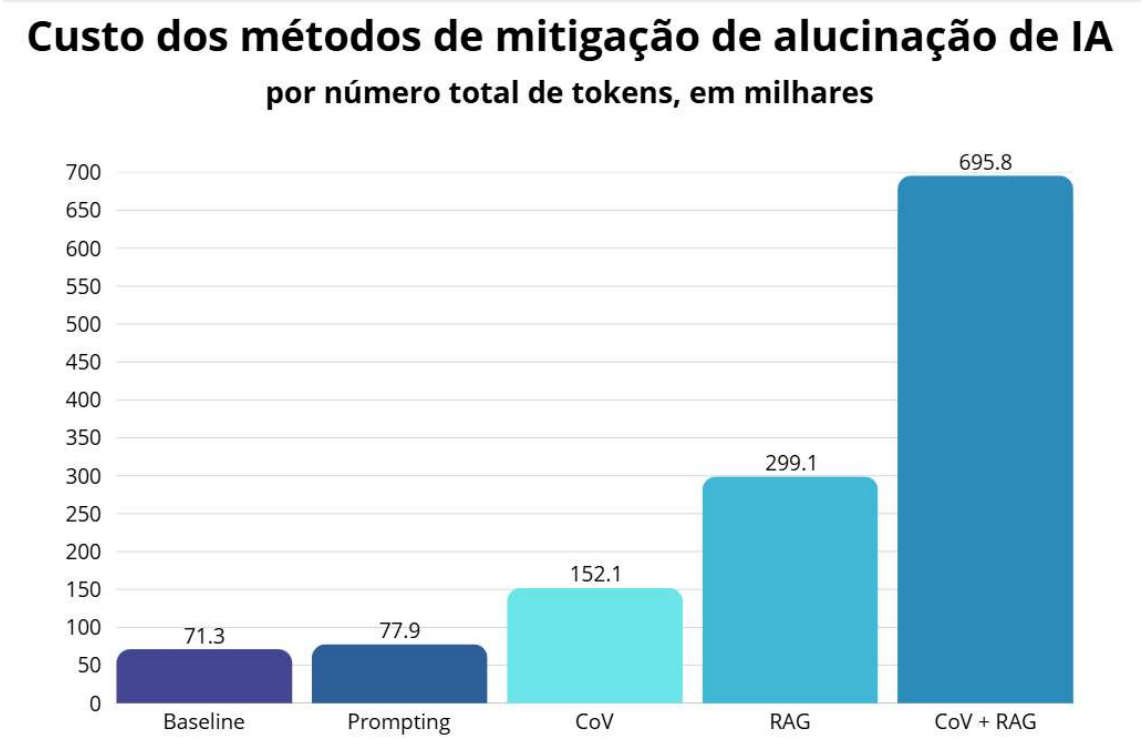
O modelo Baseline obteve a pontuação mais baixa, com 58%, reforçando os riscos apontados por Amer (2024), que observou que modelos sem mecanismos de controle apresentam maior incidência de respostas inventadas ou não verificáveis,

o que também foi perceptível especialmente nas questões ambíguas e adversariais do experimento.

O método CoV apresentou melhora limitada, confirmando o alerta de Saleh et al. (2025) de que processos de verificação desconectados de fontes confiáveis têm eficácia restrita, pois os modelos podem apenas reformular erros sem validá-los. Esse comportamento também foi mapeado nos resultados aqui obtidos.

Já o RAG, que consultava uma base de conhecimento específica, alcançou desempenho mais elevado (74%), o que converge com a literatura que caracteriza a integração entre geração e recuperação de informações como uma das estratégias mais promissoras para reduzir alucinações factuais (Sun et al., 2024; Abdelghafour, 2024). Contudo, como também indicado por Weber, Cardoso e Conter (2024), a eficácia dessa técnica depende da qualidade e contextualização da base consultada, o que explica alguns resultados parcialmente corretos observados no estudo.

Figura 9 - Gráfico demonstrando o custo dos métodos testados, em milhares de tokens.



Fonte: Do Autor.

Quando comparados os consumos de tokens entre os métodos, o impacto do custo se torna evidente. O Prompting consumiu 77.930 tokens, apresentando a melhor relação custo-benefício, com apenas 9,6 tokens utilizados por ponto de desempenho. O Baseline, mesmo com baixo consumo (71.276 tokens), obteve confiabilidade reduzida, o que o torna eficiente apenas em cenários de baixíssimo custo e tolerância a erro. Já o RAG, embora tenha sido o modelo com melhor fundamentação, ou seja, com uma base de dados mais informativa, exigiu 299.109 tokens, quase quatro vezes o consumo do Prompting, além de obter uma acurácia 7 pontos abaixo do mesmo.

O método CoV isolado, com 152.083 tokens, mostrou-se pouco vantajoso, dobrando o consumo em relação ao Baseline, mas entregando ganho mínimo de confiabilidade (3,5%). Já o CoV + RAG, com 695.808 tokens, obteve apenas 4,5 pontos percentuais de melhoria em relação ao RAG, tornando-se o método mais caro por ponto de acurácia conquistado. Em termos práticos, esse resultado evidencia que o aumento da complexidade do pipeline, adicionando camadas de verificação e consulta, gera retornos decrescentes quando aplicado a perguntas de complexidade moderada.

Essa análise reforça o trade-off entre custo computacional e confiabilidade. Em situações reais de atendimento, nas quais milhares de consultas são processadas diariamente, a diferença de consumo pode representar custos substanciais. Por exemplo, se considerarmos o preço médio atual por mil tokens, o custo de execução do CoV + RAG seria aproximadamente nove vezes maior que o Prompting, com ganho de apenas quatro pontos percentuais em precisão. Esse dado sugere que, para tarefas de suporte simples e previsíveis, Prompting oferece o melhor equilíbrio entre custo e confiabilidade.

Em contrapartida, em ambientes onde a fidedignidade da informação é crítica, como suporte jurídico, técnico avançado ou áreas de risco regulatório, o uso de estratégias como RAG ou RAG + CoV pode ser justificado, já que a prioridade passa a ser a minimização de erros factuais, mesmo com custos mais elevados.

Em síntese, os resultados obtidos demonstram que não existe uma solução universal para o problema da alucinação. Cada método apresenta vantagens e limitações específicas, e a escolha depende diretamente do contexto de aplicação e

da relação custo–benefício desejada. No cenário estudado, o Prompting apresentou o melhor desempenho geral, unindo simplicidade, baixo consumo e alta precisão. Já o RAG e o CoV + RAG mostraram potencial superior em ambientes que demandam respostas mais complexas e factuais, desde que o custo computacional elevado seja justificável

4 CONCLUSÃO

O presente estudo teve como objetivo analisar diferentes estratégias de mitigação de alucinações em chatbots baseados em inteligência artificial, utilizando o modelo GPT-5 mini da OpenAI integrado à plataforma n8n. Foram testados cinco métodos: Baseline, Prompting, Chain of Verification (CoV), Retrieval-Augmented Generation (RAG) e CoV + RAG, a fim de avaliar o impacto de cada abordagem sobre a confiabilidade das respostas e o custo computacional de execução.

Os resultados mostraram que o Prompting apresentou o melhor equilíbrio entre custo e precisão, alcançando 81% de acerto com consumo moderado de tokens. O Baseline, embora simples, obteve desempenho razoável em perguntas objetivas, mas demonstrou vulnerabilidade a alucinações em questões ambíguas. O CoV isolado teve impacto limitado, reforçando que a verificação sem base factual não garante maior confiabilidade. Já o RAG elevou a precisão das respostas para 74%, demonstrando ganhos significativos quando o modelo possui acesso a informações reais, ainda que com maior custo. O método CoV + RAG obteve o segundo melhor desempenho (78,5%), mas com consumo de tokens muito superior, mostrando que a melhoria adicional não compensou o custo elevado.

De forma geral, verificou-se que a complexidade do método nem sempre se traduz em melhor desempenho. Estratégias simples, como a engenharia de prompts, podem ser suficientes em atendimentos com perguntas diretas e repetitivas, enquanto abordagens mais elaboradas, como o RAG e o CoV + RAG, são mais indicadas para contextos em que a veracidade da informação é prioritária.

Conclui-se que a escolha do método ideal deve considerar o nível de confiabilidade exigido e o custo disponível. O Prompting se mostrou a melhor alternativa prática para cenários de suporte técnico de baixa complexidade, enquanto

o RAG e suas variações são mais adequados para aplicações críticas, nas quais a precisão justifica o investimento maior.

Como perspectivas futuras, recomenda-se a realização de novos experimentos utilizando uma base de dados ampliada, com um número maior de perguntas e maior diversidade temática, incluindo questões mais complexas e subjetivas, de modo a avaliar a capacidade dos modelos em lidar com contextos ambíguos e interpretações abertas. Além disso, propõe-se a condução de um estudo complementar com a aplicação da técnica de *fine-tuning* supervisionado, permitindo o treinamento manual do modelo com dados específicos do domínio de atendimento ao usuário. Essa abordagem possibilitaria medições mais precisas de desempenho e confiabilidade, aproximando os testes das condições observadas em ambientes corporativos reais.

REFERÊNCIAS

ABDELGHAFOUR, M. A. M. **Hallucination Mitigation Techniques in Large Language Models**. Egyptian Journal of Computing and Information Science, 2024. Disponível em: https://ijicis.journals.ekb.eg/article_406701.html. Acesso em: 02 nov. 2025.

AMER, P. **Is your AI hallucinating? New approach can tell when chatbots make things up**. Science, 2024. Disponível em: https://www.science.org/content/article/is-your-ai-hallucinating-new-approach-can-tell-when-chatbots-make-things-up. Acesso em: 09 nov. 2025.

ANTONINI, L. et al. **Applications of Large Language Models to Customer.** ScienceDirect, 2025. Disponível em: https://www.sciencedirect.com/science/article/pii/S2590123025032335. Acesso em: 09 nov. 2025.

AZAGA, H. **Effectiveness of artificial intelligence chatbots for customer service**. Middle Georgia State University, 2024. Disponível em: https://comp.mga.edu/static/media/doctoralpapers/2024_Azaga_0909150837.pdf. Acesso em: 02 nov. 2025.

BUEHLER, T. **AI in Customer Service: Revolutionizing Digital Retail**. American Public Universtiy, 2024. Disponível em: https://www.apu.apus.edu/area-of-study/business-and-management/resources/ai-in-customer-service/. Acesso em: 09 nov. 2025.

GHOSH, S. et al. **The Role of AI Enabled Chatbots in Omnichannel Customer Service**. Journal of Economics, Business & Research, v. 8, n. 4, p. 1184-1198, 2023. Disponível em: https://www.researchgate.net/publication/381096092. Acesso em: 02 nov. 2025.

HUA, H. et al. **Ophthalmology. Study on AI hallucinations**. JAMA Network, 2024. Disponível em: https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2807442. Acesso em: 09 nov. 2025.

LIN, S. et al. **TruthfulQA: Measuring How Models Mimic Human Falsehoods**. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics – Volume 1: Long Papers, 2022. Association for Computational Linguistics. Disponível em: https://www.researchgate.net/publication/361063493. Acesso em: 11 nov. 2025.

MEDURI, S. **Revolutionizing Customer Service: The Impact of Large Language Models on Chatbot Performance**. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2024. Disponível em: https://www.researchgate.net/publication/385150863. Acesso em: 04 mar. 2025.

PESSANHA, G. et al. **Large Language Models (LLMs): A systematic study in Administration, Management and Business**. Revista de Administração Mackenzie, 2024. Disponível em: https://www.scielo.br/j/ram/a/GCJHtgxWBNVv7kB73pKbm5m/. Acesso em: 09 nov. 2025.

RAZA, M. et al. **Industrial applications of large language models**. Scientific Reports, v. 15, p. 13755, 2025. Disponível em: https://doi.org/10.1038/s41598-025-98483-1. Acesso em: 09 nov. 2025.

SALEH, Y. et al. **Evaluating Large Language Models: A Comprehensive Survey**. Frontiers in Computer Science, 2025. Disponível em: https://doi.org/10.3389/fcomp.2025.1523699. Acesso em: 09 nov. 2025.

SHAHBANDI, M. **AI Chatbots vs. Human Customer Service: Impact on Consumer Satisfaction and Purchase Decisions**. International Business & Entrepreneurship Studies, v. 7, n. 3, p. 36, 2025. Disponível em: https://www.researchgate.net/publication/391944683. Acesso em: 02 nov. 2025.

SHEIKH, H. **Artificial Intelligence: Definition and Background. Artificial Intelligence and Legal Regulations**. Springer, 2023. Disponível em: https://link.springer.com/chapter/10.1007/978-3-031-21448-6_2. Acesso em: 09 nov. 2025.

Sun, Y. et al. **AI Hallucination: Towards a Comprehensive Classification of Distorted Information in AIGC**. Humanities and Social Sciences Communications, 2024. Disponível em: https://www.nature.com/articles/s41599-024-03811-x. Acesso em: 25 mar. 2025.

UZOKA, A. et al. **Leveraging AI-Powered chatbots to enhance customer service efficiency and future opportunities in automated support**. Computer Science & IT Research Journal, v. 5, n. 10, p. 2485-2510, 2024. Disponível em: https://www.researchgate.net/publication/385230161. Acesso em: 02 nov. 2025.

WEBER, G. O.; CARDOSO, L. S.; CONTER, A. S. **Chatbot e Suporte ao Cliente: um Estudo para Desenvolvimento de um Protótipo em uma Empresa de Software**. Anais do Latinoware – SBC, 2024. Disponível em: https://sol.sbc.org.br/index.php/latinoware/article/view/31566. Acesso em: 25 mar. 2025.

XIAO, Y. **Do you actually need an LLM? Rethinking language**. Artificial Intelligence Review, 2025. Disponível em: https://link.springer.com/article/10.1007/s10462-025-11308-5. Acesso em: 09 nov. 2025.

ZHANG, S. et al. **When AI Chatbots Help People Act More Human**. Harvard Business School Working Knowledge, 2025. Disponível em: https://www.library.hbs.edu/working-knowledge/when-ai-chatbots-help-people-be-more-human. Acesso em: 09 nov. 2025.