ATENDIMENTO AO CLIENTE BASEADO EM INTELIGÊNCIA ARTIFICIAL COM EXECUÇÃO CONTROLADA: IMPLEMENTAÇÃO E AVALIAÇÃO

Caio Salgado Nepomuceno
Adriana Paula Borges

Resumo

O atendimento ao cliente representa um dos principais custos operacionais em empresas com elevado volume de interações. Embora os modelos de linguagem de grande escala (LLMs) ampliem as possibilidades de automação, limitações relacionadas à geração de informações sem base factual ainda restringem sua adoção em ambientes críticos. Este trabalho apresenta o CoffAI, um sistema de atendimento baseado em execução controlada por ferramentas. A proposta restringe a atuação do modelo à identificação da intenção do usuário e ao acionamento de funções previamente definidas. Os resultados obtidos em cinquenta interações de teste indicaram precisão na seleção de ferramentas e baixa ocorrência de alucinações, sugerindo que a abordagem contribui para aumentar a confiabilidade e a previsibilidade de sistemas conversacionais aplicados ao atendimento ao cliente.

Palavras-chave: *Atendimento ao cliente. Automação. Chatbots. Inteligência artificial. Redução de custos.*

Abstract
Customer service represents one of the main operational costs for companies with a high volume of interactions. Although large-scale language models (LLMs) expand the possibilities for automation, limitations related to the generation of unfounded information still restrict their adoption in critical environments. This paper presents CoffAI, a customer service system based on tool-controlled execution. The proposal restricts the model's role to identifying the user's intent and triggering predefined functions. Results obtained from fifty test interactions indicated accuracy in tool selection and a low incidence of hallucinations, suggesting that the approach contributes to increasing the reliability and predictability of conversational systems applied to customer service.

Keywords*: Artificial intelligence. Automation. Chatbots. Cost reduction. Customer service.*

1 Introdução

O atendimento ao cliente é uma atividade estratégica para organizações que mantêm relacionamento contínuo com seus consumidores. Além de influenciar a satisfação e a fidelização, representa um dos principais custos operacionais de empresas de médio e grande porte. Em organizações com elevado volume de interações, parcela significativa dos recursos é destinada ao suporte, tornando sua eficiência um fator relevante para a competitividade (BUTTLE; MAKLAN, 2019).

Grande parte dessas interações consiste em demandas repetitivas, como esclarecimento de dúvidas, consultas de informações e execução de procedimentos padronizados, características que favorecem sua automação. Nesse cenário, os modelos de linguagem de grande escala (Large Language Models – LLMs) ampliaram as possibilidades de desenvolvimento de sistemas conversacionais capazes de compreender e responder solicitações em linguagem natural.

Apesar dos avanços, a utilização de LLMs em ambientes corporativos ainda enfrenta desafios relacionados à confiabilidade das respostas. Embora produzam textos coerentes, esses modelos podem gerar informações incorretas ou inexistentes, fenômeno conhecido como alucinação (JI et al., 2023). No atendimento ao cliente, esse comportamento pode resultar em retrabalho, aumento de reclamações e perda de credibilidade, exigindo mecanismos adicionais para garantir maior precisão das informações fornecidas.

Entre as estratégias para mitigar esse problema, destaca-se a execução controlada por ferramentas (tools). Nessa abordagem, o modelo atua como identificador de intenções e orquestrador de funções especializadas, enquanto as respostas são obtidas por meio de consultas a bases de dados, sistemas corporativos ou serviços externos. Dessa forma, as informações tornam-se verificáveis e rastreáveis, reduzindo a geração de conteúdo sem respaldo factual.

Nesse contexto, foi desenvolvido o CoffAI, um sistema de atendimento automatizado baseado em execução controlada por ferramentas. O sistema realiza consultas a dados estruturados, cálculo de prazos de entrega e localização de unidades físicas, encaminhando ao atendimento humano solicitações fora do escopo definido. Ao restringir deliberadamente as ações disponíveis ao modelo de linguagem, busca-se aumentar a previsibilidade e a confiabilidade das respostas.

A questão de pesquisa deste trabalho é: em que medida a execução controlada por ferramentas pode aumentar a confiabilidade de sistemas de atendimento automatizado sem comprometer sua capacidade de resolução? Parte-se da hipótese de que a restrição do modelo a um conjunto previamente definido de ferramentas reduz a ocorrência de alucinações, aumenta a previsibilidade das respostas e diminui a necessidade de intervenção humana em demandas rotineiras. Para investigar essa hipótese, o sistema CoffAI foi desenvolvido e avaliado, conforme apresentado nas seções seguintes.

2 Referencial Teórico e Trabalhos Correlatos

2.1 Chatbots e Automação no Atendimento

A automação do atendimento ao cliente tem sido amplamente adotada por organizações que buscam aumentar a eficiência operacional e reduzir custos de suporte (ADAMOPOULOU; MOUSSIADES, 2020). Nesse contexto, os chatbots consolidaram-se como uma das principais tecnologias para automatizar interações recorrentes, como esclarecimento de dúvidas, direcionamento de usuários e execução de processos padronizados.

Entretanto, a eficácia desses sistemas não depende apenas da correção das respostas. Segundo Shum, He e Li (2018), a qualidade percebida está relacionada à naturalidade da interação, à consistência das respostas e à capacidade de compreender diferentes formas de expressão dos usuários. Limitações nesses aspectos podem gerar frustração e comprometer os benefícios da automação.

Historicamente, os chatbots foram baseados em classificação de intenções e fluxos conversacionais predefinidos. Nessa abordagem, cada mensagem é associada a uma intenção específica que determina a ação executada pelo sistema (BOCKLISCH et al., 2017). Embora amplamente utilizada, sua eficácia depende da cobertura das intenções previstas durante o desenvolvimento, tornando o sistema vulnerável a solicitações não contempladas.

Os avanços em Processamento de Linguagem Natural (PLN) reduziram essas limitações ao ampliar a capacidade de interpretação das intenções independentemente da formulação utilizada pelo usuário. Esse progresso viabilizou a incorporação de modelos de linguagem de grande escala aos sistemas

conversacionais, aumentando sua flexibilidade e adaptação a diferentes contextos de interação.

2.2 Modelos de linguagem e limitações

Os modelos de linguagem de grande escala (Large Language Models – LLMs) representam um dos avanços mais relevantes da Inteligência Artificial aplicada à linguagem natural. Treinados em grandes volumes de dados textuais, esses modelos demonstram elevada capacidade de compreender contexto, gerar respostas coerentes e executar tarefas linguísticas complexas (BROWN et al., 2020).

O aumento da escala dos modelos tem sido acompanhado pelo surgimento de capacidades emergentes, como raciocínio em múltiplas etapas, síntese de informações e resolução de problemas complexos. Wei et al. (2022) demonstram que modelos com bilhões de parâmetros podem desenvolver mecanismos intermediários de raciocínio que melhoram o desempenho em tarefas de inferência sequencial.

Apesar desses avanços, os LLMs não possuem mecanismos internos de validação factual. Como consequência, podem gerar informações incorretas ou inexistentes apresentadas de forma plausível, fenômeno conhecido como alucinação. Esse comportamento decorre da natureza probabilística da geração textual, baseada na previsão da sequência de palavras mais provável, e não na verificação da veracidade das informações produzidas (JI et al., 2023).

No atendimento ao cliente, essa limitação é particularmente crítica, pois as respostas fornecidas frequentemente subsidiam decisões dos usuários. Informações incorretas sobre produtos, serviços, prazos ou políticas empresariais podem gerar impactos operacionais, financeiros e reputacionais, tornando a confiabilidade um requisito essencial para a adoção de LLMs em ambientes corporativos (AMODEI et al., 2016).

2.3 Abordagens para Mitigação de Alucinação

A necessidade de aumentar a confiabilidade dos modelos de linguagem impulsionou o desenvolvimento de estratégias para mitigação de alucinações. Entre as mais difundidas está o Retrieval-Augmented Generation (RAG), que combina geração de linguagem com mecanismos de recuperação de informação (LEWIS et al.,

2020). Nessa abordagem, documentos ou bases de conhecimento externas são consultados antes da geração da resposta, permitindo fundamentar o conteúdo em fontes atualizadas e verificáveis.

Embora reduza a dependência da memória paramétrica do modelo, o RAG mantém a síntese sob responsabilidade do LLM, preservando a possibilidade de interpretações inadequadas ou de geração de conteúdo não suportado pelas evidências recuperadas.

Outra linha de pesquisa explora a integração de ferramentas externas ao processo de geração. O Toolformer (SCHICK et al., 2023) demonstrou que modelos de linguagem podem acionar recursos como calculadoras, mecanismos de busca e APIs para executar tarefas com maior precisão. De forma complementar, o framework ReAct (YAO et al., 2023) propõe a alternância entre raciocínio e execução de ações, utilizando resultados intermediários para orientar etapas subsequentes do processo decisório.

Embora reduzam a dependência da geração puramente textual, essas abordagens ainda preservam graus de liberdade para inferência do modelo, mantendo a possibilidade de respostas inconsistentes em cenários que exigem comportamento determinístico.

A efetividade dessas estratégias é normalmente avaliada por métricas de confiabilidade, como precisão factual (factual accuracy), precisão na seleção de ferramentas (tool selection accuracy) e taxa de resolução sem intervenção humana (containment rate) (DERIU et al., 2021). Essas métricas permitem mensurar tanto a qualidade das respostas quanto o impacto operacional da automação em ambientes reais.

2.4 Trabalhos Correlatos

A literatura recente apresenta uma tendência crescente de integração entre modelos de linguagem e mecanismos externos de recuperação de informação ou execução de ações. Trabalhos baseados em RAG buscam aumentar a precisão das respostas por meio do acesso a fontes externas de conhecimento (LEWIS et al., 2020), enquanto abordagens fundamentadas em agentes ampliam a capacidade operacional dos modelos por meio da utilização de ferramentas especializadas (SCHICK et al., 2023; YAO et al., 2023).

Bocklisch et al. (2024) argumentam que arquiteturas que combinam modelos de linguagem e execução determinística de lógica de negócio apresentam maior escalabilidade do que sistemas baseados exclusivamente em classificação de intenções. Da mesma forma, iniciativas como o AutoGPT (SIGNIFICANT GRAVITAS, 2023) demonstram o potencial de agentes baseados em LLMs para coordenar múltiplas ações de forma autônoma em tarefas complexas.

Entre os trabalhos analisados, o ReAct (YAO et al., 2023) apresenta a maior proximidade conceitual com a proposta desenvolvida neste estudo. Contudo, enquanto o ReAct combina raciocínio e execução de ferramentas em ciclos iterativos, o CoffAI adota uma estratégia mais restritiva, limitando a atuação do modelo à identificação da intenção do usuário e à seleção da ferramenta apropriada. Nesse modelo, a resposta é derivada diretamente do retorno da ferramenta executada, reduzindo a dependência de inferências produzidas pelo modelo.

A principal diferença em relação às abordagens RAG reside no tratamento da etapa de síntese. Enquanto sistemas baseados em recuperação aumentada ainda dependem da capacidade do modelo para interpretar e sintetizar as informações recuperadas, o CoffAI busca minimizar essa etapa, privilegiando respostas fundamentadas em dados estruturados e retornos determinísticos de ferramentas específicas.

Além disso, observa-se que grande parte dos trabalhos correlatos concentra sua avaliação em métricas de desempenho linguístico ou precisão factual, dedicando menor atenção aos impactos operacionais da automação em ambientes reais de atendimento. Aspectos como taxa efetiva de absorção de demandas, comportamento diante de solicitações fora do escopo e necessidade de intervenção humana permanecem relativamente pouco explorados. É precisamente nessa lacuna que o presente trabalho se insere, ao investigar o potencial da execução controlada por ferramentas como mecanismo para aumentar a confiabilidade e a previsibilidade de sistemas de atendimento automatizado.

3 MATERIAIS E MÉTODOS

O CoffAI foi desenvolvido como um sistema de atendimento ao cliente baseado em execução controlada por ferramentas, com o objetivo de investigar o impacto dessa abordagem na confiabilidade de sistemas conversacionais apoiados por

modelos de linguagem. Diferentemente de arquiteturas que permitem a geração livre de respostas, o sistema restringe a atuação do modelo à identificação da intenção do usuário e à seleção da ferramenta mais adequada para atender à solicitação recebida.

A proposta fundamenta-se na premissa de que a utilização de fontes estruturadas e mecanismos determinísticos de obtenção de informações pode reduzir a dependência de inferências produzidas pelo modelo de linguagem e, consequentemente, diminuir a ocorrência de respostas sem respaldo factual. Dessa forma, o modelo atua como um componente de orquestração, enquanto a obtenção das informações é delegada a funções especializadas previamente definidas.

As ferramentas disponibilizadas foram selecionadas a partir de cenários representativos de atendimento ao cliente no domínio de aplicação adotado neste estudo, contemplando consultas de receitas de café, estimativas de prazo de entrega, localização de unidades físicas e suporte ao programa de associados. Para situações que excedem o escopo das funcionalidades implementadas, o sistema realiza o encaminhamento da solicitação para atendimento humano, preservando a continuidade do processo de atendimento.

3.2 Arquitetura do Sistema

O CoffAI foi implementado como uma aplicação web distribuída, organizada em arquitetura cliente-servidor e estruturada em camadas com responsabilidades bem definidas. Essa separação visa favorecer a manutenção do sistema, a escalabilidade da solução e o isolamento entre os componentes responsáveis pela interface do usuário e pelo processamento das requisições.

A camada de apresentação foi desenvolvida utilizando o framework Next.js em conjunto com a biblioteca React. Essa camada é responsável pela renderização da interface conversacional e pela comunicação com os serviços disponibilizados pelo backend. O gerenciamento das interações com o modelo de linguagem é realizado por meio do pacote @ai-sdk/react, que fornece mecanismos para envio e recebimento assíncrono de mensagens.

A camada de processamento foi implementada utilizando o framework AdonisJS, executado sobre a plataforma Node.js e desenvolvido em TypeScript. Essa camada concentra a lógica de negócio da aplicação, incluindo o processamento das

mensagens recebidas, a identificação de intenções, a seleção de ferramentas e a construção das respostas encaminhadas ao usuário.

Os dados estruturados utilizados pelas ferramentas são armazenados em um banco de dados relacional PostgreSQL, responsável por centralizar informações referentes às receitas de café, unidades físicas, produtos e demais dados necessários ao funcionamento do sistema. Para operações relacionadas ao cálculo de frete, foi utilizada uma API intermediária responsável por abstrair a comunicação com os serviços de entrega, reduzindo a complexidade de integração e desacoplando a aplicação de dependências específicas de provedores externos.

3.3 Definição das Ferramentas

As funcionalidades do sistema foram implementadas por meio de ferramentas independentes, concebidas como módulos especializados responsáveis pela execução de tarefas específicas. Cada ferramenta opera de forma isolada e pode ser acionada conforme a intenção identificada na mensagem do usuário.

As ferramentas implementadas são:

- Consulta de receitas de café: recupera receitas armazenadas no banco de dados a partir dos parâmetros fornecidos pelo usuário;
- Estimativa de prazo de entrega: consulta serviços externos para calcular prazos estimados de entrega de acordo com o destino informado;
- Localização de unidades físicas: recupera informações cadastrais referentes às unidades disponíveis;
- Suporte ao programa de associados: fornece informações relacionadas ao programa de fidelidade com base em dados estruturados;
- Encaminhamento para atendimento humano (fallback): realiza a transferência da solicitação para um operador quando nenhuma das ferramentas disponíveis é capaz de atender adequadamente à demanda apresentada.

Nessa arquitetura, o modelo de linguagem não é responsável pela obtenção direta das informações apresentadas ao usuário. Sua função restringe-se à interpretação da solicitação e à seleção da ferramenta apropriada para cada contexto. A resposta final é construída a partir dos dados retornados pelas ferramentas

executadas, reduzindo significativamente a dependência de inferências produzidas pelo modelo e aumentando a rastreabilidade das informações fornecidas.

Embora essa estratégia contribua para a redução da ocorrência de alucinações, não se afirma que tal comportamento seja completamente eliminado. Como demonstrado posteriormente nos resultados experimentais, ainda podem ocorrer situações excepcionais associadas ao gerenciamento de contexto conversacional ou a comportamentos emergentes do modelo.

3.4 Fluxo de Funcionamento

O fluxo operacional do CoffAI inicia-se com o envio de uma mensagem pelo usuário por meio da interface conversacional. A mensagem é encaminhada ao backend, onde é processada pelo modelo de linguagem com o objetivo de identificar a intenção subjacente à solicitação e determinar qual ferramenta deverá ser acionada.

Após a seleção da ferramenta apropriada, o sistema executa a função correspondente. Dependendo da natureza da solicitação, essa execução pode envolver consultas ao banco de dados, acesso a serviços externos ou recuperação de informações previamente estruturadas. Os resultados obtidos são então utilizados para compor a resposta apresentada ao usuário.

A arquitetura foi projetada para minimizar a geração livre de conteúdo, priorizando a utilização de informações provenientes de fontes controladas. Dessa forma, busca-se aumentar a consistência, a previsibilidade e a verificabilidade das respostas produzidas pelo sistema.

Quando nenhuma das ferramentas disponíveis é capaz de atender à solicitação recebida, o mecanismo de fallback é acionado e a interação é encaminhada para atendimento humano. Nesse processo, as informações coletadas durante a conversa são preservadas e disponibilizadas ao operador responsável, reduzindo a necessidade de repetição de informações por parte do usuário e contribuindo para a continuidade do atendimento.

Essa estratégia permite que limitações de cobertura funcional sejam tratadas por transferência para um agente humano, em vez de depender exclusivamente da capacidade de inferência do modelo de linguagem, reduzindo os riscos associados à geração de respostas inadequadas em situações fora do escopo previsto.

4 Resultados e Discussão

4.1 Cenários de Uso

A avaliação experimental do CoffAI foi estruturada a partir de cinco cenários representativos de demandas frequentemente observadas em sistemas de atendimento ao cliente:

- Consulta de receitas de café;
- Estimativa de prazo de entrega;
- Localização de unidades físicas;
- Suporte ao programa de associados;
- Solicitações fora do escopo, com encaminhamento para atendimento humano.

A definição desses cenários teve como objetivo avaliar a capacidade da arquitetura proposta de operar em diferentes tipos de interação, contemplando tanto situações diretamente atendidas pelas ferramentas implementadas quanto casos que exigem mecanismos de tratamento alternativo. Dessa forma, a avaliação não se restringe à verificação do funcionamento individual das ferramentas, mas busca analisar o comportamento do sistema como um todo diante de diferentes condições operacionais.

A seleção dos cenários foi orientada por dois critérios metodológicos. O primeiro corresponde à representatividade funcional, segundo a qual cada cenário representa uma categoria distinta de operação executada pelo sistema. Enquanto a consulta de receitas e a localização de unidades envolvem recuperação de dados estruturados, a estimativa de prazo de entrega requer integração com um serviço externo. O suporte ao programa de associados representa o acesso a informações institucionais específicas, enquanto o cenário de solicitações fora do escopo avalia a capacidade do sistema de reconhecer seus próprios limites operacionais. Em conjunto, esses cenários abrangem diferentes formas de interação comumente encontradas em ambientes reais de atendimento automatizado.

O segundo critério refere-se à verificabilidade dos resultados. Em todos os cenários considerados, a correção da resposta pode ser determinada objetivamente por meio da comparação com dados previamente conhecidos ou com informações retornadas pelas fontes consultadas. Essa característica permite avaliar o desempenho do sistema com base em evidências observáveis, reduzindo a influência

de julgamentos subjetivos relacionados à qualidade linguística das respostas e direcionando a análise para aspectos centrais da proposta, como a seleção adequada de ferramentas, a precisão das informações fornecidas e a confiabilidade do fluxo de atendimento.

A inclusão do cenário de solicitações fora do escopo possui relevância particular para os objetivos deste estudo. Avaliar um sistema conversacional apenas em situações para as quais ele foi explicitamente projetado tende a superestimar seu desempenho, uma vez que não considera as condições de incerteza e variabilidade presentes em ambientes reais de utilização. Conforme destacado por Deriu et al. (2021), a avaliação de sistemas de diálogo deve incluir situações limítrofes e casos não previstos, permitindo observar como a aplicação responde quando confrontada com demandas que extrapolam suas capacidades operacionais.

Nesse contexto, o mecanismo de encaminhamento para atendimento humano não foi tratado como um recurso de exceção, mas como um componente essencial da arquitetura proposta. A capacidade de identificar solicitações não atendidas pelas ferramentas disponíveis e direcioná-las adequadamente para um operador humano constitui parte fundamental da estratégia de confiabilidade adotada pelo CoffAI. Consequentemente, o desempenho desse mecanismo é avaliado com o mesmo rigor aplicado às funcionalidades automatizadas, uma vez que sua eficácia influencia diretamente a qualidade global do atendimento oferecido pelo sistema.

A partir desses cenários, torna-se possível analisar não apenas a precisão das respostas produzidas, mas também a efetividade da abordagem de execução controlada por ferramentas na redução de comportamentos indesejados, na manutenção da rastreabilidade das informações e na condução adequada de situações que demandam intervenção humana.

4.2 Demonstração do Sistema

Com o objetivo de ilustrar o funcionamento do CoffAI e evidenciar o comportamento da arquitetura proposta em diferentes situações de uso, esta seção apresenta exemplos de interações realizadas nos cenários descritos anteriormente. As figuras documentam etapas representativas do fluxo operacional do sistema, desde a interação inicial com o usuário até a execução das ferramentas especializadas e o tratamento de solicitações fora do escopo de automação.

As Figuras 1 a 6 apresentam evidências qualitativas do funcionamento do CoffAI nos diferentes cenários avaliados. A Figura 1 mostra a interface principal da aplicação, enquanto a Figura 2 ilustra uma interação em linguagem natural, evidenciando a capacidade do modelo de interpretar solicitações sem necessidade de comandos estruturados.

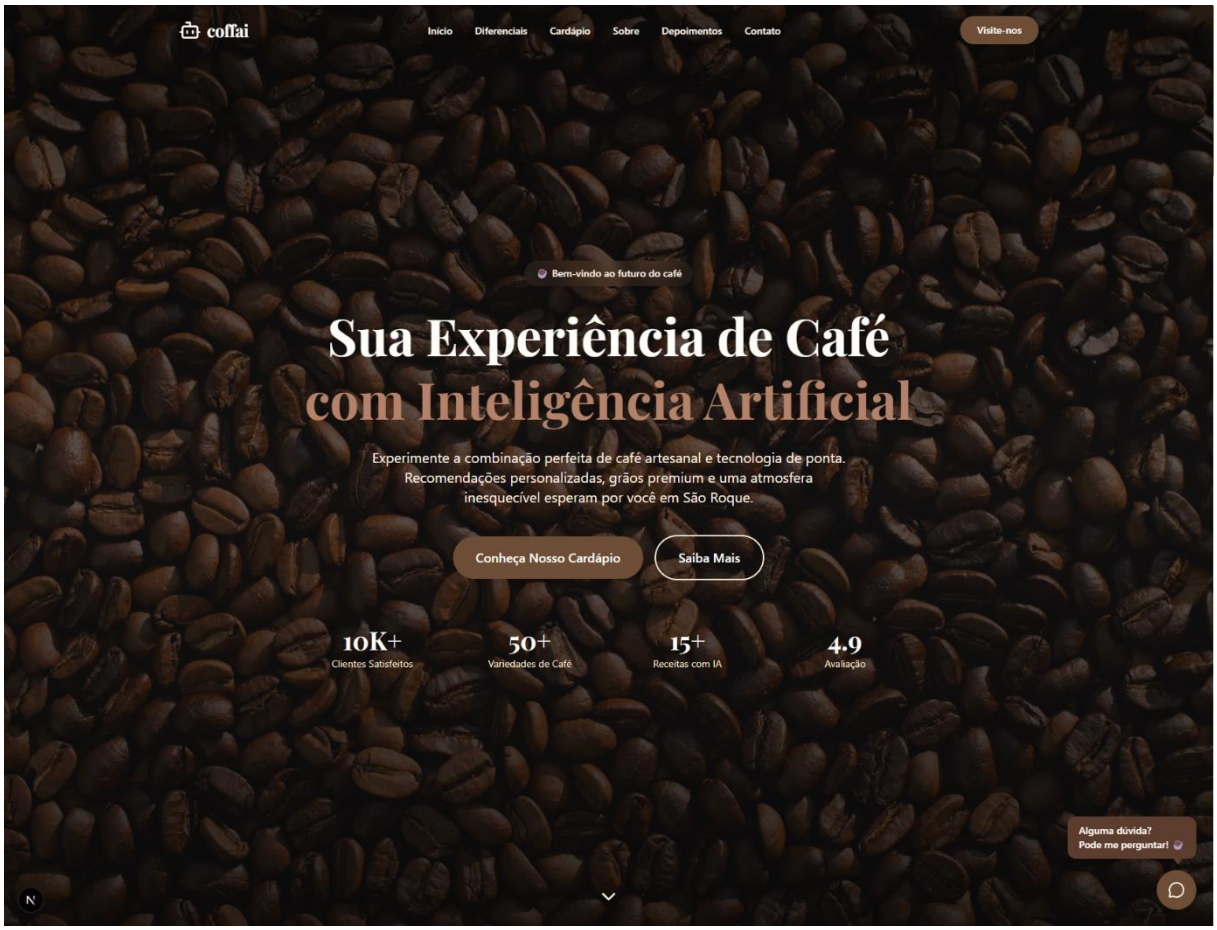

Figura 1 – Página inicial da aplicação CoffAI com interface conversacional integrada
Fonte: Elaborado pelo autor (2026).

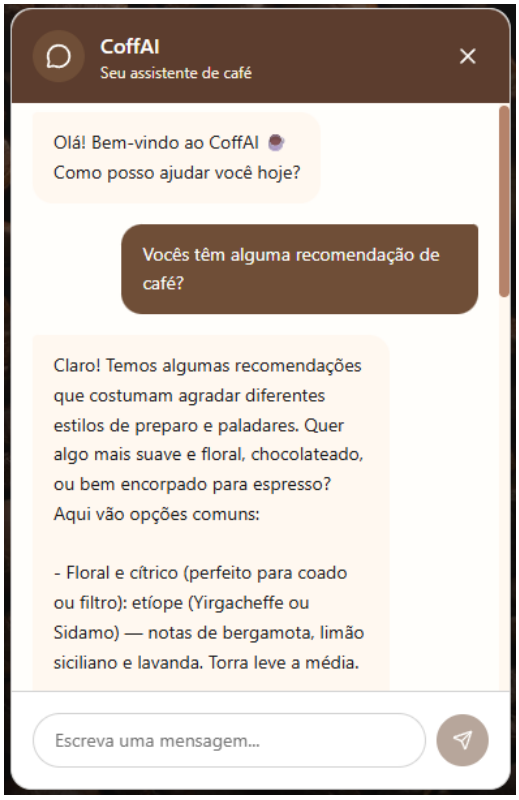

Figura 2 – Interação inicial com o chatbot CoffAI
Fonte: Elaborado pelo autor (2026).

As Figuras 3, 4 e 5 demonstram a execução das ferramentas implementadas para recuperação de dados estruturados, consulta a serviços externos e obtenção de informações cadastrais. Em todos os casos, o modelo identifica a intenção do usuário e aciona a ferramenta apropriada, evidenciando a separação entre interpretação da solicitação e obtenção da resposta, princípio central da arquitetura proposta.

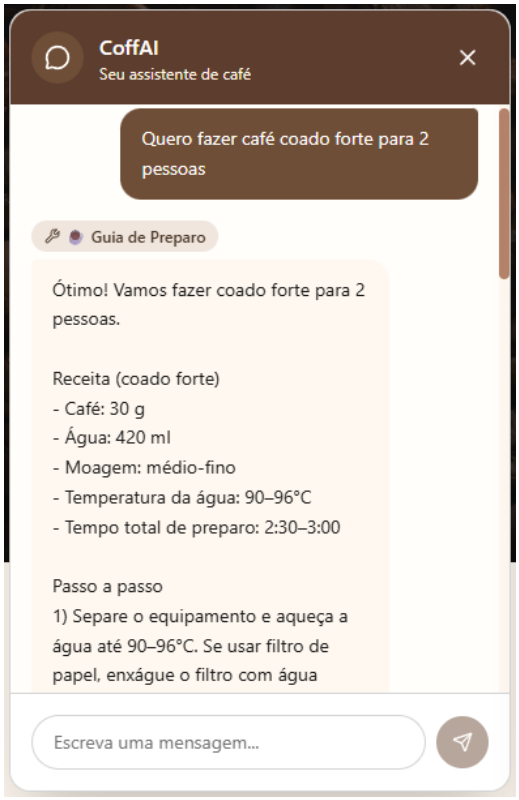

Figura 3 – Geração de receita de café a partir de execução de ferramenta
Fonte: Elaborado pelo autor (2026).

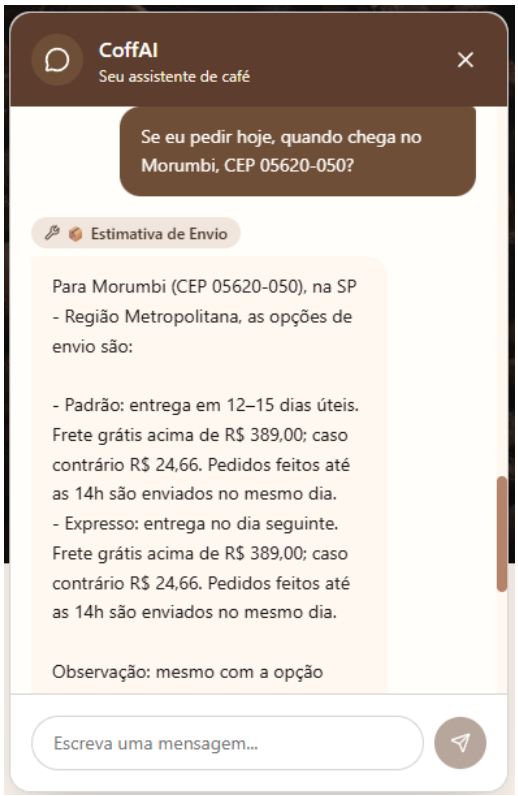

Figura 4 – Estimativa de prazo de entrega com uso de API externa
Fonte: Elaborado pelo autor (2026).

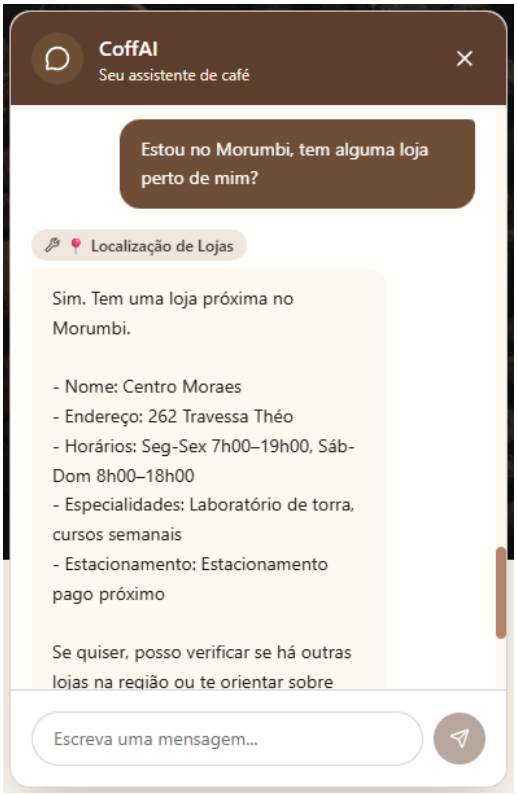

Figura 5 – Consulta de localização de unidades utilizando dados estruturados
Fonte: Elaborado pelo autor (2026).

Um aspecto relevante é a rastreabilidade das respostas. O sistema registra explicitamente a ferramenta utilizada em cada interação, permitindo verificar a origem das informações fornecidas. Essa característica amplia a transparência e a auditabilidade do atendimento, diferenciando a abordagem de sistemas baseados predominantemente em geração livre de texto.

A Figura 6 ilustra o acionamento do mecanismo de fallback quando a solicitação não pode ser atendida pelas ferramentas disponíveis. Nesses casos, a interação é encaminhada ao atendimento humano com preservação do histórico da conversa e das informações coletadas, permitindo a continuidade do atendimento sem que o usuário precise repetir dados previamente informados. Esse comportamento evidencia que o CoffAI prioriza a confiabilidade do atendimento ao reconhecer explicitamente seus limites de cobertura funcional.

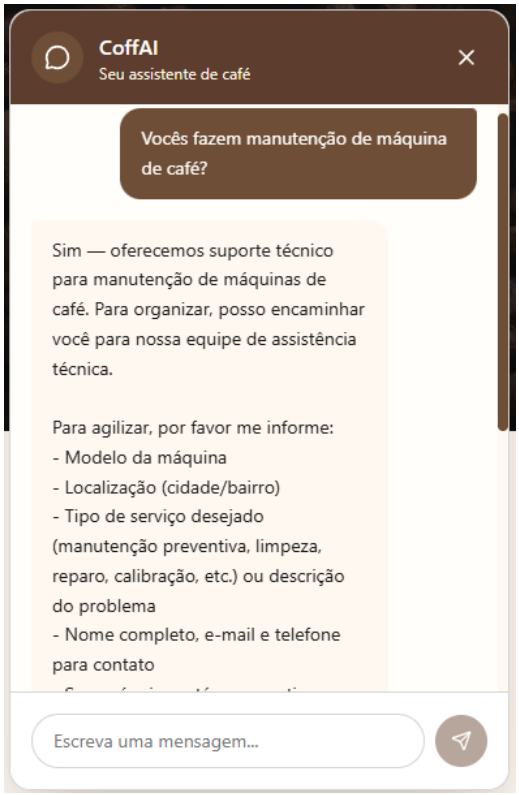

Figura 6 – Encaminhamento estruturado para atendimento humano
Fonte: Elaborado pelo autor (2026).

Embora as figuras permitam observar qualitativamente o funcionamento da arquitetura, a validação da proposta exige métricas objetivas de desempenho. Por esse motivo, foi realizada uma avaliação quantitativa baseada em interações controladas, apresentada na seção seguinte.

4.3 Avaliação Quantitativa

Para avaliar o desempenho do CoffAI, foi elaborado um conjunto de 50 mensagens de teste, distribuídas igualmente entre os cinco cenários definidos na Seção 4.1, totalizando 10 interações por cenário.

Com o objetivo de avaliar a robustez da identificação de intenções, os testes incluíram diferentes formas de formulação, abrangendo linguagem formal, coloquial, expressões ambíguas e casos de borda. Dessa forma, buscou-se reproduzir variações frequentemente observadas em ambientes reais de atendimento.

As métricas adotadas seguiram as recomendações de Deriu et al. (2021) para avaliação de sistemas conversacionais, contemplando: (i) precisão na seleção da ferramenta apropriada (*tool selection accuracy*), (ii) precisão da resposta final

(*response accuracy*), (iii) taxa de resolução sem intervenção humana (*containment rate*), (iv) adequação do mecanismo de *fallback* e (v) ocorrência de alucinações. Os resultados são apresentados nas Tabelas 1 e 2 e discutidos nas seções subsequentes.

Tabela 1 – Métricas gerais de desempenho

Métrica	Resultado
Taxa de ferramenta correta	90,0% (45/50)
Taxa de resposta correta	88,0% (44/50)
Taxa de resolução automática	76,0% (38/50)
Taxa de fallback adequado	66,7% (8/12)
Taxa de alucinação detectada	2,0% (1/50)

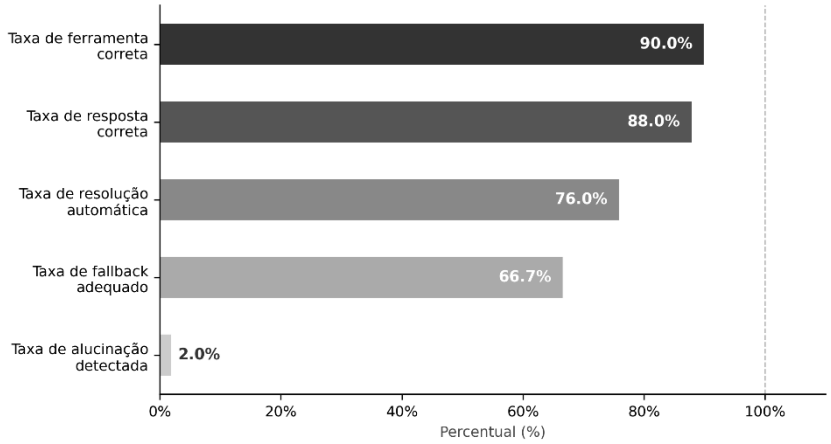

Figura 7 – Métricas gerais de desempenho do CoffAI
Fonte: Elaborado pelo autor (2026).

Tabela 2 – Desempenho por cenário

Cenário	Ferramenta correta	Resposta correta	Fallbacks
Receitas de café	90% (9/10)	100% (10/10)	1
Prazo de entrega	100% (10/10)	100% (10/10)	0
Localização de unidades	90% (9/10)	70% (7/10)	2
Programa de associados	90% (9/10)	90% (9/10)	1
Fora do escopo / fallback	80% (8/10)	80% (8/10)	8

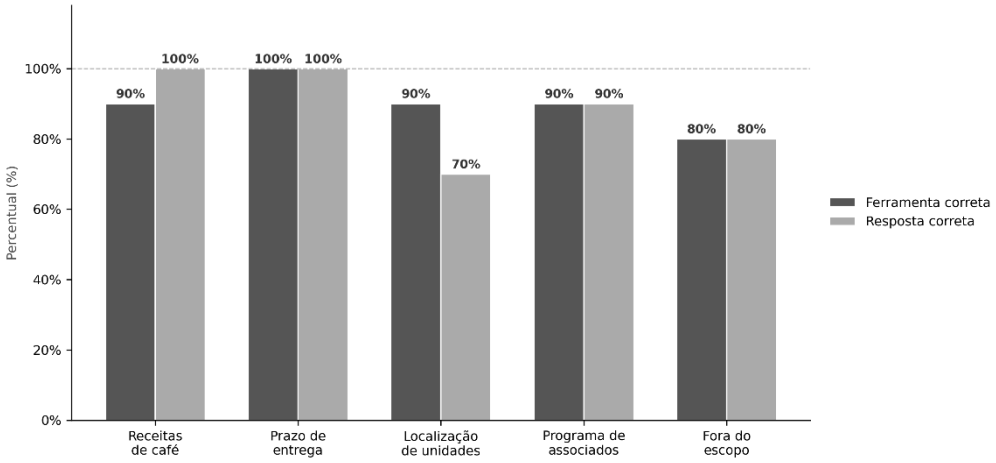

Figura 8 – Desempenho por cenário: precisão da ferramenta e sua resposta
Fonte: Elaborado pelo autor (2026).

4.3.1 Análise dos resultados

O cenário de prazo de entrega apresentou o desempenho mais consistente: 100% de precisão na seleção de ferramenta e 100% de precisão na resposta final. A integração com a API de frete mostrou-se estável e o modelo demonstrou capacidade de reconhecer a intenção associada a esse cenário mesmo em formulações ambíguas. Na mensagem 14 ("Preciso receber antes do fim de semana, é possível?"), o sistema solicitou a região antes de responder e confirmou a viabilidade do frete expresso. Na mensagem 18, além de informar a disponibilidade, orientou o usuário a fornecer o destino para obtenção de uma estimativa mais precisa.

O cenário de localização de unidades registrou a maior discrepância entre precisão de ferramenta (90%) e precisão de resposta (70%). Essa diferença não decorre de alucinação — o modelo não gerou endereços sem base factual — mas de ausência de dados: consultas sobre horário de funcionamento acionaram a ferramenta corretamente, porém a informação solicitada não estava disponível no banco de dados. Esse resultado evidencia que a qualidade das respostas em sistemas baseados em ferramentas depende não apenas da arquitetura, mas da completude dos dados acessados pelas funções executadas.

Dos 12 encaminhamentos para atendimento humano registrados, 8 ocorreram nos casos de solicitações fora do escopo — comportamento esperado e consistente com o funcionamento previsto do mecanismo de fallback. Os 4 encaminhamentos restantes ocorreram em cenários cobertos pelas ferramentas disponíveis, em situações de formulação ambígua nas quais o modelo optou pela transferência em vez de inferir a intenção do usuário. Esse comportamento conservador é coerente com

o objetivo de minimizar respostas incorretas: em sistemas de atendimento ao cliente, o custo operacional de um encaminhamento desnecessário é inferior ao custo de uma resposta inadequada (BUTTLE; MAKLAN, 2019).

A única ocorrência de alucinação foi registrada na mensagem 50. Após uma sequência de interações, o usuário realizou uma pergunta de acompanhamento e o sistema produziu uma resposta sem acionar ferramenta, gerando uma explicação incorreta sobre modalidades de pagamento. A análise indica que o evento não decorreu de geração livre por ausência de dado, mas de perda de contexto conversacional em interação de múltiplas trocas. Embora pontual, o caso evidencia uma limitação da abordagem em sessões conversacionais mais extensas, aspecto que será retomado na seção 4.4.

Por fim, dois casos apresentaram comportamento distinto do fluxo controlado previsto. Nas mensagens 30 ("Vocês têm loja em Portugal?") e 33 ("Tem desconto pra quem é membro?"), nenhuma ferramenta foi acionada, mas as respostas produzidas foram consideradas adequadas. O modelo utilizou o contexto disponível para chegar a conclusões corretas sem passar pelo fluxo de execução controlada. Esse comportamento emergente não compromete a validade da abordagem, mas indica que a fronteira entre execução controlada e raciocínio contextual nem sempre corresponde ao que o sistema foi configurado para produzir — aspecto relevante para trabalhos futuros que busquem maior controle sobre o comportamento do modelo em casos limítrofes.

4.4 Limitações observadas

Os resultados demonstram desempenho satisfatório do sistema no escopo proposto, mas revelam três limitações relevantes.

A primeira é estrutural e decorre da própria arquitetura adotada. A restrição à execução de ferramentas predefinidas, responsável por reduzir alucinações, faz com que solicitações fora desse conjunto sejam encaminhadas ao atendimento humano. Nos testes, esse comportamento ocorreu em quatro casos de encaminhamento conservador, nos quais demandas potencialmente atendíveis foram transferidas devido à ambiguidade da formulação. Trata-se de uma limitação inerente à execução controlada: o mesmo mecanismo que aumenta a confiabilidade reduz a cobertura do sistema (SHUM; HE; LI, 2018).

A segunda limitação foi observada no cenário de localização de unidades, em que a precisão da ferramenta (90%) superou a precisão da resposta (70%). A diferença ocorreu porque o banco de dados não continha informações sobre horário de funcionamento. Embora a ferramenta correta tenha sido acionada, os dados necessários para a resposta estavam ausentes, evidenciando a dependência desses sistemas da qualidade e completude das bases consultadas.

A terceira limitação corresponde à única alucinação registrada, ocorrida em uma interação de múltiplas trocas. Nesse caso, o sistema perdeu o contexto conversacional e respondeu sem acionar ferramenta. Conforme destacam Shum, He e Li (2018), a gestão de contexto em diálogos prolongados permanece um dos principais desafios dos sistemas conversacionais, exigindo mecanismos adequados de rastreamento do histórico. Nos testes com mensagens isoladas, o comportamento foi consistente, indicando que o risco aumenta com a extensão da conversa.

Essas limitações — estrutural, de dados e de contexto conversacional — não comprometem a validade da abordagem proposta, mas definem seus limites operacionais e indicam prioridades para trabalhos futuros.

Referências

ADAMOPOULOU, Eleni; MOUSSIADES, Lefteris. An overview of chatbot technology. In: MAGLOGIANNIS, Ilias; ILIADIS, Lazaros; PIMENIDIS, Elias (Org.). **Artificial Intelligence Applications and Innovations**. Cham: Springer, 2020. p. 373–383.

AMODEI, Dario et al. Concrete problems in AI safety. **arXiv preprint**, arXiv:1606.06565, 2016. Disponível em: https://arxiv.org/abs/1606.06565. Acesso em: maio 2026.

BOCKLISCH, Tom et al. Rasa: Open Source Language Understanding and Dialogue Management. **arXiv preprint**, arXiv:1712.05181, 2017. Disponível em: https://arxiv.org/abs/1712.05181. Acesso em: maio 2026.
BOCKLISCH, Tom et al. Conversational AI with Language Models. **arXiv preprint**, arXiv:2402.12234, 2024. Disponível em: https://arxiv.org/abs/2402.12234. Acesso em: maio 2026.

BROWN, Tom B. et al. Language models are few-shot learners. In: **Advances in Neural Information Processing Systems**, v. 33, p. 1877–1901, 2020.

BUTTLE, Francis; MAKLAN, Stan. **Customer Relationship Management: Concepts and Technologies**. 4. ed. Londres: Routledge, 2019.

CHOWDHERY, Aakanksha et al. PaLM: Scaling language modeling with pathways. **arXiv preprint**, arXiv:2204.02311, 2022. Disponível em: https://arxiv.org/abs/2204.02311. Acesso em: maio 2026.

DERIU, Jan et al. Survey on evaluation methods for dialogue systems. **Artificial Intelligence Review**, v. 54, n. 1, p. 755–810, 2021.

JI, Zhenzhong et al. Survey of hallucination in natural language generation. **ACM Computing Surveys**, v. 55, n. 12, art. 248, p. 1–38, 2023.

LEWIS, Patrick et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: **Advances in Neural Information Processing Systems**, v. 33, p. 9459–9474, 2020.

SCHICK, Timo et al. Toolformer: language models can teach themselves to use tools. In: **Advances in Neural Information Processing Systems**, v. 36, p. 68539–68551, 2023.

SHUM, Heung-Yeung; HE, Xiao-dong; LI, Di. From Eliza to XiaoIce: challenges and opportunities with social chatbots. **Frontiers of Information Technology & Electronic Engineering**, v. 19, n. 1, p. 10–26, 2018.

SIGNIFICANT GRAVITAS. **Auto-GPT: An Autonomous GPT-4 Experiment**. GitHub, 2023. Disponível em: https://github.com/Significant-Gravitas/AutoGPT. Acesso em: maio 2026.

WEI, Jason et al. Chain-of-thought prompting elicits reasoning in large language models. In: **Advances in Neural Information Processing Systems**, v. 35, p. 24824–24837, 2022.

YAO, Shunyu et al. ReAct: Synergizing reasoning and acting in language models. In: **International Conference on Learning Representations (ICLR)**, 2023. Disponível em: https://arxiv.org/abs/2210.03629. Acesso em: maio 2026.

YAO, Shunyu et al. Tree of thoughts: deliberate problem solving with large language models. In: **Advances in Neural Information Processing Systems**, v. 36, p. 11809–11822, 2023.

Orientadora: Foi utilizada a ferramenta **ChatGPT (OpenAI)** para apoio na revisão gramatical, aprimoramento da redação acadêmica e adequação do texto às normas científicas. O conteúdo técnico, a análise dos resultados e as conclusões foram elaborados e validados pelos autores.