PROJETO DE UM CHATBOT IA PARA INSIGHTS DO E-COMMERCE BRASILEIRO

John Kelvin Gonçalves de Souza[1],
Alexandre Adler Cunha de Freitas[2],
Elionai de Souza Magalhães[2],
Kevyn Phillipe Gusmão[2]

[1]Discente do curso de Engenharia de computação do Centro Universitário Multivix Vitória
[2] Docentes do Centro Universitário Multivix Vitória

RESUMO
Este projeto propõe o desenvolvimento e implementação de um chatbot com Inteligência Artificial para o comércio eletrônico brasileiro, visando obter insights valiosos e impulsionar a eficiência operacional. A pesquisa, que se baseia no crescimento da IA no país e na necessidade de personalização do comércio eletrônico, desenvolveu um modelo com interface de chat e treinamento em dados sintéticos. A avaliação em ambiente de testes revelou que a decomposição de perguntas e o roteamento adaptativo com APIs foram mais eficientes que o modelo RAG com Self-RAG, especialmente em cenários complexos. Recomenda-se melhorias na compreensão de dados temporais complexos e a continuação da exploração de abordagens de roteamento inteligente e integração de APIs para fornecer recomendações estratégicas mais relevantes e acessíveis. O estudo demonstra o potencial da IA para o desenvolvimento de chatbots inteligentes que podem auxiliar lojistas na tomada de decisões estratégicas no e-commerce, abrindo caminho para um futuro mais eficiente e personalizado para o comércio eletrônico brasileiro.
PALAVRAS-CHAVE
Agente; Chatbot; E-commerce; Inteligência Artificial; Recomendação.

ABSTRACT
This project proposes the development and implementation of an Artificial Intelligence chatbot for Brazilian e-commerce, aimed at gaining valuable insights and enhancing operational efficiency. Based on the growth of AI in the country and the need for personalization in e-commerce, the research developed a model with a chat interface and training on synthetic data. Testing in a simulated environment showed that question decomposition and adaptive routing with APIs were more effective than the RAG model with Self-RAG, particularly in complex scenarios. Improvements in understanding complex temporal data and further exploration of intelligent routing approaches and API integration are recommended to provide more relevant and accessible strategic recommendations. The study demonstrates the potential of AI in developing intelligent chatbots that can assist retailers in making strategic decisions in e-commerce, paving the way for a more efficient and personalized future for Brazilian e-commerce.
KEYWORDS
Agent; Chatbot; E-commerce; Artificial Intelligence; Recommendation.

INTRODUÇÃO

Com o progresso da inteligência artificial, assistentes virtuais têm surgido com muita frequência para auxiliar pessoas em diversas áreas, destacando-se pela capacidade de processamento de dados e aprendizado contínuo. A ferramenta ChatGPT, lançada no final de 2022, é um exemplo notável dessa tendência,

contribuindo significativamente para a popularização da IA tanto no meio acadêmico quanto na indústria (LIU *et al.*, 2023).

Proposto por Turing (1950), o Teste de Turing é uma referência no campo do desenvolvimento da inteligência artificial (IA), oferecendo uma definição operacional de inteligência. Para ser aprovado no teste, um programa deve demonstrar habilidades cruciais, incluindo processamento de linguagem natural (PLN), representação de conhecimento, raciocínio automatizado e aprendizado de máquina (NORVIG, 2013).

O processamento de linguagem natural é essencial para o funcionamento de agentes conversacionais. Através dele, os programas interpretam texto ou fala de entrada e produzem respostas úteis e adequadas, possibilitando a realização de conversas de forma alternada (BENGFORT; BILBRO; OJEDA, 2018). O aprendizado de máquina complementa essa capacidade, permitindo que os sistemas se adaptem com base em novos dados e lidem com problemas complexos que envolvem grandes quantidades de informações (GÉRON, 2021).

À medida que a tecnologia avança, a ampla implementação de arquiteturas de aprendizagem profunda para diversas tarefas de processamento de linguagem natural se mostra vantajosa. Essa tecnologia permite que os sistemas aprendam a utilizar conhecimentos prévios para compreender informações e capturar os aspectos semânticos e sintáticos relevantes, proporcionando um processamento da linguagem mais eficiente e preciso (DENG; LIU, 2018). Esta evolução tem contribuído significativamente para o aprimoramento de sistemas de inteligência artificial no contexto da linguagem natural.

Sendo essenciais para gerenciar interações dos usuários, essas tecnologias empregadas em chatbots de IA são de altíssimo valor para fornecer experiências personalizadas de forma interativa, contribuindo assim para a satisfação do usuário e, por conseguinte, proporcionando benefícios significativos para os negócios que as adotam.

Em um cenário marcado pela rápida migração para o e-commerce durante a pandemia, impulsionada pelo fechamento das lojas físicas e pelo consequente aumento exponencial das vendas online conforme relatado por E-Commerce Brasil (2021), muitos lojistas se viram diante do desafio de se adaptar a um novo mercado de forma rápida e eficiente.

Diante desse contexto e da liderança do Brasil no uso de inteligência artificial na América Latina (SAS, 2022), este projeto se propôs a auxiliar os lojistas, tanto os experientes quanto os que ingressaram recentemente no e-commerce a obter insights valiosos, melhorar a tomada de decisões estratégicas e impulsionar a eficiência operacional.

O foco principal do projeto foi investigar o funcionamento de um chatbot equipado com IA e sua aplicação no cenário do comércio eletrônico brasileiro. Para isso, foi desenvolvida uma interface interativa, utilizando dados sintéticos provenientes do ambiente online, permitindo a geração de informações relevantes tanto em uma perspectiva global do e-commerce quanto de forma específica para atender às necessidades dos lojistas em suas vendas.

1. REFERENCIAL TEÓRICO

Esta seção apresenta o referencial teórico utilizado para embasar o trabalho proposto. Inicialmente, são abordados os fundamentos das tecnologias de chatbots inteligentes, destacando os conceitos essenciais e as principais tecnologias que possibilitam o desenvolvimento desses sistemas. Em seguida, são exploradas as tecnologias envolvidas na criação de um agente de inteligência artificial, como algoritmos de aprendizado profundo e arquiteturas de redes neurais. Por fim, a seção detalha como um agente com foco em dados do comércio eletrônico brasileiro pode ser construído, analisando os desafios e soluções específicos desse contexto.

1.1 Fundamentos das Tecnologias de Chatbots Inteligentes

“ELIZA é um programa que opera dentro do sistema de compartilhamento de tempo MAC do MIT, que torna possíveis certos tipos de conversação em linguagem natural entre homem e computador” (WEIZENBAUM, 1966, p. 01, tradução nossa).

Sendo o precursor da interação natural entre humano e computador, ELIZA foi essencial para o entendimento e avanço da área de chatbots, influenciando significativamente na evolução dos chatbots modernos. Com o objetivo de permitir uma comunicação mais estruturada e eficiente, seu desenvolvimento envolveu a formulação e aplicação de um conjunto predefinido de regras de funcionamento conforme é ilustrado na Figura 1.

Figura 1 – Interface de usuário do ELIZA

Fonte: Akyon (2018).

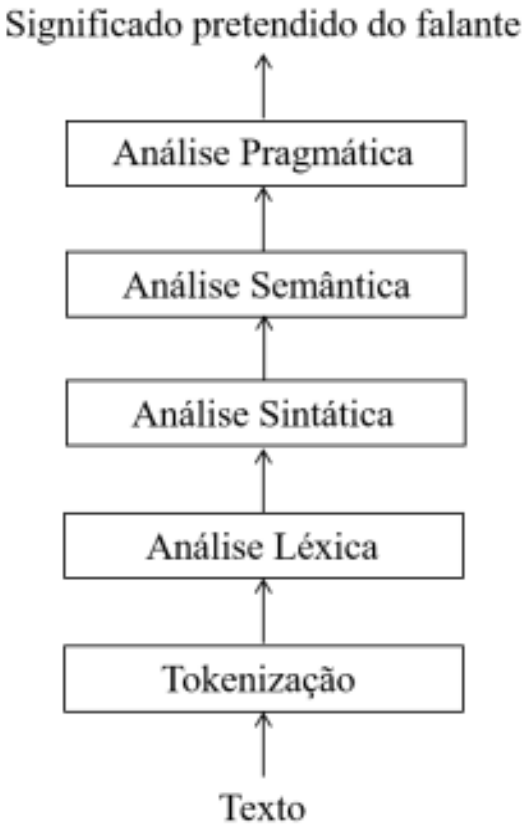
Segundo Schlicht (2016), os chatbots podem funcionar de duas formas: por regras ou por aprendizado de máquina (AM), sendo o primeiro mais limitado, podendo responder consultas simples com respostas predefinidas. Dessa forma os chatbots com inteligência artificial vem ganhando popularidade atualmente, sendo capazes de aprender e se adaptar às necessidades dos usuários.

Conforme Martins *et al*. (2020), até meados dos anos 1980, os algoritmos eram principalmente criados com base em conjuntos de regras manuais, contudo, ao final da década, houve uma revolução no processamento de linguagem natural com a introdução de algoritmos de aprendizado automático. Como base para diversas tecnologias atuais, o aprendizado de máquina é uma subárea da IA bem utilizada para solucionar a crescente complexidade dos problemas, e lidar com o aumento da velocidade e volume de dados gerados por diferentes setores (FACELI *et al*., 2021).

Para que os chatbots possam entender e interpretar a linguagem humana de forma natural, é preciso utilizar ferramentas capazes de permitir essa interação. Uma dessas é o PLN, que representa um subcampo da inteligência artificial dedicado a capacitar computadores a processar, interpretar e gerar linguagem humana de forma natural (BLETE; CAELEN, 2023).

No processamento de linguagem natural, a análise da linguagem é dividida em estágios que abrangem várias tarefas, como classificação de texto, geração de texto, resposta a perguntas e tradução. Esses estágios refletem questões teóricas linguísticas, como sintaxe, semântica e pragmatismo, podendo ser decompostos em mais etapas conforme representado na Figura 2 (DAMERAU; INDURKHYA, 2010).

Figura 2 – Etapas de avaliação na execução de processamento de linguagem natural



Fonte: Damerau e Indurkhya (2010).

Para Bird, Klein e Loper (2009), outros modelos, como redes neurais multiníveis, são muito mais obscuros, mesmo sendo possível obter insights é preciso muito mais trabalho. No entanto, com o avanço em poder computacional, otimizações nas taxas de aprendizagem e o surgimento de bibliotecas para auxiliar no desenvolvimento, as redes neurais estão rapidamente se tornando referência no campo do processamento de linguagem natural (BENGFORT; BILBRO; OJEDA, 2018).

1.2 Tecnologias de um Agente de Inteligência Artificial

Para (Deng; Liu, 2018), o aprendizado profundo se origina das redes neurais profundas, que são caracterizadas por várias camadas dispostas em cascata. Essas redes são inspiradas em sistemas neurais biológicos, buscando imitar a complexidade e a capacidade de aprendizado do cérebro humano. Conhecidas como segmentos dentro do campo de aprendizado profundo, modelos como redes neurais recorrentes (RNNs), que podem ser ajustadas para lidar com o processamento de sequências extremamente longas de dados, e redes de memória de curto prazo (LSTMs), uma variação das RNNs que melhora a capacidade de aprendizado a longo prazo utilizando auto-loops, têm se destacado (GOODFELLOW; BENGIO; COURVILLE, 2016).

Além disso, as redes neurais convolucionais (CNNs ou ConvNets) têm sido tremendamente bem-sucedidas em aplicações práticas, como no reconhecimento de imagens e no processamento de vídeo. Outro modelo importante são as redes generativas adversárias (GANs), que têm revolucionado a geração de dados sintéticos, oferecendo novas possibilidades para a criação de imagens, textos e outros tipos de dados. A popularidade desses modelos aumentou significativamente nos últimos anos, conforme apontam (BENGFORT; BILBRO; OJEDA, 2018).

De acordo com Zhang *et al.* (2023), os LSTMs são caracterizados por possuírem três tipos de portas: entrada, esquecimento e saída, que desempenham a função de regular o fluxo de informações. Na camada oculta do LSTM, ocorre a geração tanto do estado oculto quanto do estado interno da célula de memória, sendo que somente o estado oculto é transmitido para a próxima camada, enquanto o estado interno permanece interno. Essa arquitetura específica contribui significativamente para mitigar problemas de gradientes desaparecendo e explodindo, permitindo que as redes aprendam com dados mais complexos e de grande extensão de forma mais produtiva e precisa.

Adicionalmente, os LSTMs se destacam como um fator importante em diversas aplicações de inteligência artificial, como reconhecimento de voz, tradução automática e processamento de linguagem natural, devido à sua capacidade de aprender sequências longas e analisar o contexto. Essa eficácia é evidenciada pela equação de atualização e pela dinâmica de retropropagação atrativa, como ressaltado por (KARPATHY, 2015).

O cenário dos modelos de linguagem neural testemunhou um progresso notável nos últimos anos, impulsionado por avanços arquitetônicos e no poder computacional. Em 2017, foi introduzido o Transformer, uma arquitetura inovadora que dispensa mecanismos recorrentes e se baseia unicamente na atenção para capturar dependências globais entre entrada e saída, permitindo uma paralelização significativa e um desempenho superior em relação aos modelos recorrentes pré-existentes (VASWANI *et al.*, 2017).

O Transformer segue uma estrutura codificador-decodificador, composta por múltiplas camadas de blocos de atenção empilhados. Cada bloco codificador engloba uma camada de autoatenção, seguida por normalização, conexão residual e uma camada de alimentação direta e os blocos decodificadores replicam a estrutura dos codificadores, incorporando adicionalmente uma camada de atenção cruzada para conectar as representações do codificador ao decodificador (YOU; SUN; IYYER, 2020).

Um aspecto fundamental dos Transformers reside na atenção cruzada e na autoatenção, presentes também em modelos de linguagem de grande escala (LLMs). Estes mecanismos, derivados do conceito de atenção, direcionam o foco para os termos mais relevantes em cada etapa da tarefa, aprimorando o processamento (BLETE; CAELEN, 2023). A autoatenção, em particular, destaca-se por sua capacidade de relacionar as palavras na entrada entre si, demonstrando grande efetividade em tarefas como geração de texto e resposta a perguntas, onde o contexto da frase é crucial para respostas precisas (VASWANI *et al.*, 2017).

Em contrapartida aos Transformers, os Generative Pre-Trained Transformer (GPT), também baseados na arquitetura geral do Transformer, apresentam uma distinção crucial: a ausência de um codificador (KUBLIK; SABOO, 2022). Essa característica elimina a necessidade de atenção cruzada, fazendo com que os GPTs dependam exclusivamente da autoatenção dentro do decodificador para gerar representações e previsões sensíveis ao contexto. Essa simplificação os torna adaptáveis a uma ampla gama de tarefas, além de reduzir significativamente o tempo de treinamento.

1.3 Agente com Foco em Dados do Comércio Eletrônico Brasileiro

No âmbito da Inteligência Artificial, os agentes de IA se destacam como componentes essenciais, integrando diversas funcionalidades para executar tarefas autônomas com alta eficiência. Conforme a documentação da Google (2024), um agente se caracteriza por metas, que descrevem o que precisa ser alcançado, e instruções, que delineiam as etapas necessárias para atingir esses objetivos. Para aprimorar o treinamento do LLM, exemplos de interações, conhecidos como comandos few-shot, são utilizados.

Aplicações que envolvem agentes de IA frequentemente empregam múltiplos agentes trabalhando em conjunto, cada um especializado em processar diferentes tipos de tarefas. Estes agentes são configurados para fornecer informações, enviar consultas a serviços externos ou realizar subtarefas. Devlin *et al*. (2018) apresentam duas estratégias principais para utilizar representações de idioma pré-treinadas em tarefas de fluxo: a abordagem baseada em recursos e a abordagem de ajuste fino. Cada uma possui características e vantagens distintas, tornando-se ferramentas valiosas para diferentes cenários.

Para Brown *et al*. (2020), o fine-tuning tem sido a abordagem mais comum nos últimos anos e envolve a atualização dos pesos de um modelo pré-treinado, treinando-o em um conjunto de dados supervisionado específico para a tarefa desejada. Normalmente, são usados milhares a centenas de milhares de exemplos rotulados. Dessa forma, utilizar o ajuste fino melhora o aprendizado em poucas tentativas, permitindo a obtenção de melhores resultados em um amplo número de tarefas (OPENAI, 2023).

Outras abordagens enriquecem as capacidades dos agentes, como a proposta de Schick *et al.* (2023) de equipar modelos de linguagem com a habilidade de utilizar várias ferramentas externas através de chamadas de Application Program Interfaces (APIs), usando tokens especiais para delimitar essas chamadas. Além disso, há abordagens que requerem interações humanas para fornecer feedback, fundamental para aprimorar o modelo e proporcionar respostas mais adequadas (CHRISTIANO *et al.*, 2017).

Como também a técnica denominada Retrieval-Augmented Generation (RAG), introduzida por Lewis *et al.* (2020); Guu *et al.* (2020), representa uma técnica inovadora que integra a geração de texto com a recuperação de informações. Essa

metodologia emprega um mecanismo de busca para recuperar documentos pertinentes e um gerador para criar respostas com base nessas informações previamente recuperadas.

Apesar da robustez e do potencial dos agentes de Inteligência Artificial, ainda existem desafios a serem superados. Um dos principais reside na possibilidade do modelo gerar respostas incorretas, mas que aparentam estar corretas, denominadas "alucinações" (MAYNEZ *et al.*, 2020). Essas falhas factuais podem comprometer a confiabilidade e a efetividade dos agentes, exigindo o desenvolvimento de técnicas para mitigá-las.

No contexto do RAG, diversos métodos foram introduzidos com o objetivo de superar suas limitações e aprimorar a qualidade da geração do LLM. Dentre elas, destacam-se a Geração Aumentada de Recuperação Autorreflexiva (Self-RAG), conforme discutido por Asai *et al.* (2023), e o Adaptive-RAG, como é apresentado por (JEONG *et al.*, 2024). A Self-RAG tem como objetivo melhorar a precisão factual do modelo, utilizando técnicas de recuperação sob demanda e autorreflexão, sem comprometer sua versatilidade.

O Adaptive-RAG propõe um novo método eficaz para recuperar informações entre várias LLMs (recuperação aumentada). Ele utiliza um classificador menor, treinado para prever o nível de dificuldade da consulta, com conjuntos de dados de treinamento coletados automaticamente sem rotulagem humana. Esse método adapta-se à complexidade específica de cada consulta, determinando o modelo mais apropriado para respondê-la (JEONG *et al.*, 2024).

No cenário atual, empresas brasileiras reconhecem a importância da análise de dados históricos e da obtenção de insights preditivos para aprimorar suas decisões estratégicas (SAS, 2022). Nesse sentido, os agentes de IA se apresentam como ferramentas valiosas para alcançar esse objetivo.

Considerando a variedade de fornecedores de LLM disponíveis no mercado, como Gemini (TEAM *et al*., 2023), Llama 2 (TOUVRON *et al.*, 2023) e GPT-4 (OPENAI *et al.*, 2023). A escolha do provedor ideal dependerá das necessidades específicas de cada tarefa, mas todos oferecem um alto grau de personalização, permitindo que as empresas adaptem os agentes aos seus fluxos de trabalho e objetivos.

2. METODOLOGIA

Este trabalho caracteriza-se como uma pesquisa aplicada, focada na busca de soluções práticas para a implementação de chatbots inteligentes. A abordagem adotada é uma combinação de métodos qualitativos e quantitativos pois fornece uma visão abrangente, permitindo uma compreensão profunda das tecnologias e ao mesmo tempo mensura o desempenho do agente em relação aos dados sintéticos do comércio eletrônico no Brasil.

No que diz respeito ao objetivo da pesquisa, adota-se uma perspectiva exploratória, buscando entender e explorar os avanços tecnológicos e sua aplicação específica em chatbots de IA. Isso está alinhado com a abordagem discutida por Wazlawick (2020), que ao examinar fenômenos em busca de anomalias, conhecidas ou não, visa estabelecer a base para uma investigação mais sistemática posteriormente. Quanto aos procedimentos, a pesquisa é de natureza bibliográfica, envolvendo uma revisão detalhada da literatura existente sobre o uso de tecnologias em chatbots.

O método de pesquisa aplicado é indutivo, partindo de observações específicas para generalizações mais amplas relacionadas ao uso de chatbots no comércio eletrônico. Visando avaliar o desempenho dos chatbots no cenário do comércio eletrônico, destacando-se como uma pesquisa que se propõe a contribuir para o entendimento e aprimoramento prático dessa tecnologia inovadora mesmo suas conclusões podendo não ser absolutas, têm alta probabilidade de serem verdadeiras caracterizando como método estatístico (MATIAS-PEREIRA, 2016).

2.1 Procedimentos Metodológicos

Para o desenvolvimento do projeto, foi realizada a geração de dados sintéticos com base em informações transacionais do comércio eletrônico, incluindo variáveis como pedidos, faturamento, ticket médio e frete médio pago, garantindo a diversidade e a representatividade dos dados observados no primeiro trimestre de 2024. Com base nesses dados coletados, foram criados dois modelos com diferentes estratégias para o desenvolvimento e implementação do chatbot.

O primeiro modelo utilizou um conjunto diversificado de conversas de demonstração, que foram organizadas em pequenos pedaços em um formato de documento. Após a organização, esses documentos passaram por uma função de

embedding, que transformou os textos em representações vetoriais. Essas representações vetoriais foram então salvas em um banco de dados vetorial. Este banco vetorial atuou como um recuperador de informações, permitindo que o chatbot respondesse às questões dos usuários de maneira eficiente e precisa. O Pinecone foi escolhido por sua capacidade de realizar buscas rápidas e precisas em grandes volumes de dados vetoriais, combinando a pesquisa vetorial com filtros de metadados familiares para obter os resultados desejados (PINECONE, 2024).

O segundo modelo se diferencia pela implementação de uma arquitetura de ferramentas e roteamento. Essa arquitetura funciona como um sistema centralizado que recebe as perguntas dos usuários e busca os dados necessários para responder às suas perguntas através de chamadas para uma API.

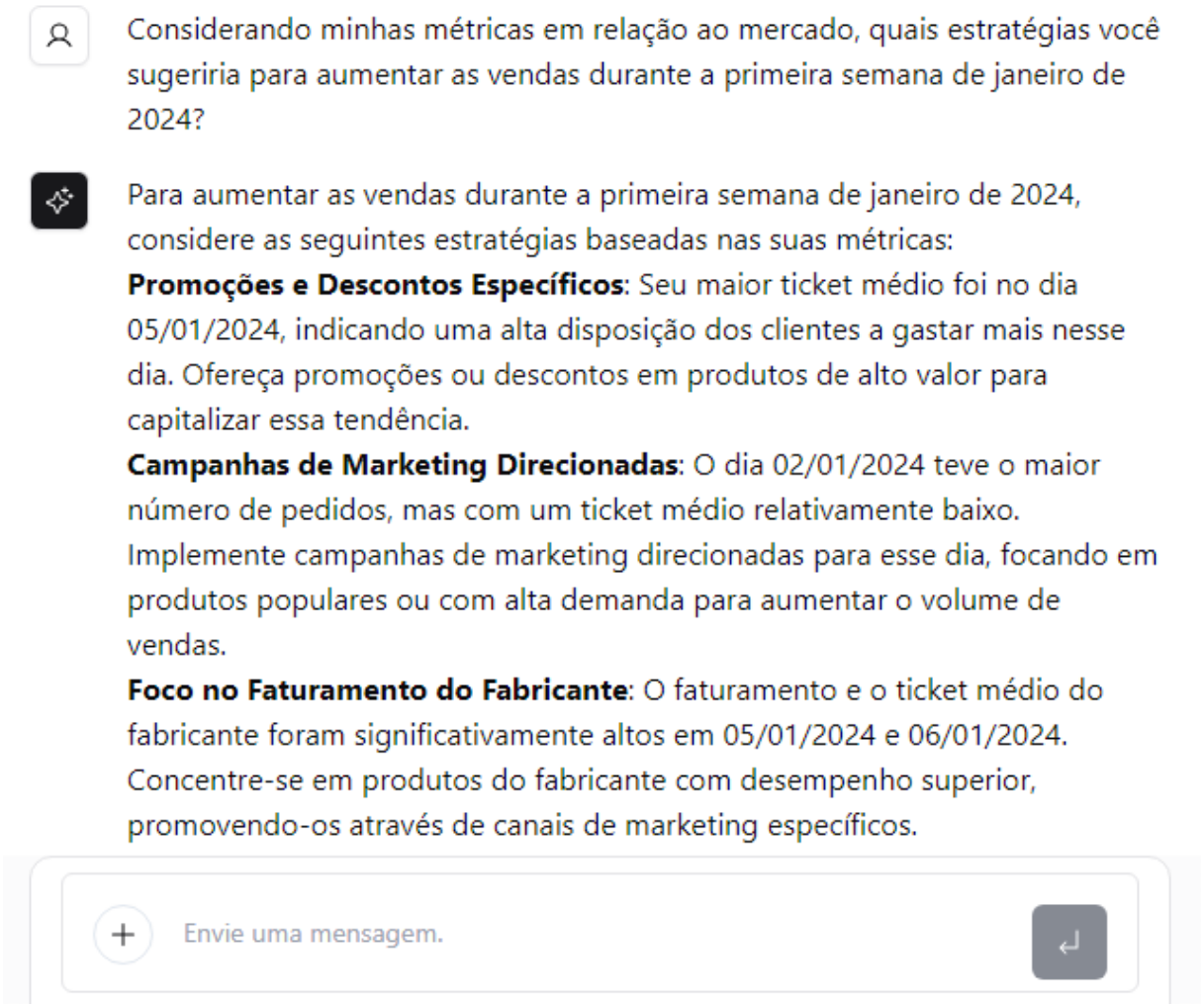
O LLM desempenhou um papel crucial nesse processo, sendo empregado para processar as solicitações dos usuários e direcioná-las para a rota correta, juntamente com os parâmetros necessários para gerar respostas mais precisas e assertivas. Também permitindo uma iteração, possibilitando que o LLM seja consultado novamente para sugerir uma recomendação caso o usuário a requisite.

Para esse funcionamento, foi utilizado os modelos de linguagem da OpenAI: o gpt-3.5-turbo-0125, que direcionou cada conversa para o caminho ideal, garantindo uma interação eficiente, e o gpt-4-0125-preview, responsável por gerar respostas precisas e personalizadas, considerando o contexto da conversa e as necessidades do cliente. Ambos os modelos são suportados pela Vercel AI SDK, uma biblioteca open source especializada em interfaces de usuários baseadas em IA (VERCEL, 2024). A integração desses modelos com a interface de conversação foi facilitada pela Vercel AI SDK, simplificando o desenvolvimento da interface, acelerando o processo de implementação e garantindo uma interação robusta entre o sistema e o usuário.

Sendo utilizada em conjunto com o framework web Next.js, amplamente adotado por grandes empresas globais (VERCEL, 2024). Essa combinação de tecnologias permitiu que a interface do chatbot fornecesse respostas rápidas e precisas, adaptando-se em tempo real às necessidades dos usuários. Além disso, a arquitetura robusta do Next.js proporcionou uma base sólida para o crescimento contínuo do chatbot, garantindo escalabilidade e facilidade na incorporação de novas funcionalidades para aprimorar a experiência do usuário.

Ao utilizar a aplicação e o modelo de ferramentas, o usuário pode ser direcionado para diferentes cenários, dependendo de suas intenções e necessidades. O sistema é projetado para identificar o objetivo do usuário a partir de suas interações. Por exemplo, se o usuário solicitar uma recomendação explícita, ele será encaminhado para o sistema de recomendação, conforme ilustrado na Figura 3. Nesse cenário, o chatbot analisará as preferências e o histórico do usuário para fornecer sugestões personalizadas, otimizando a relevância das recomendações.

Figura 3 – Interface do chatbot com resultados no formato de recomendação

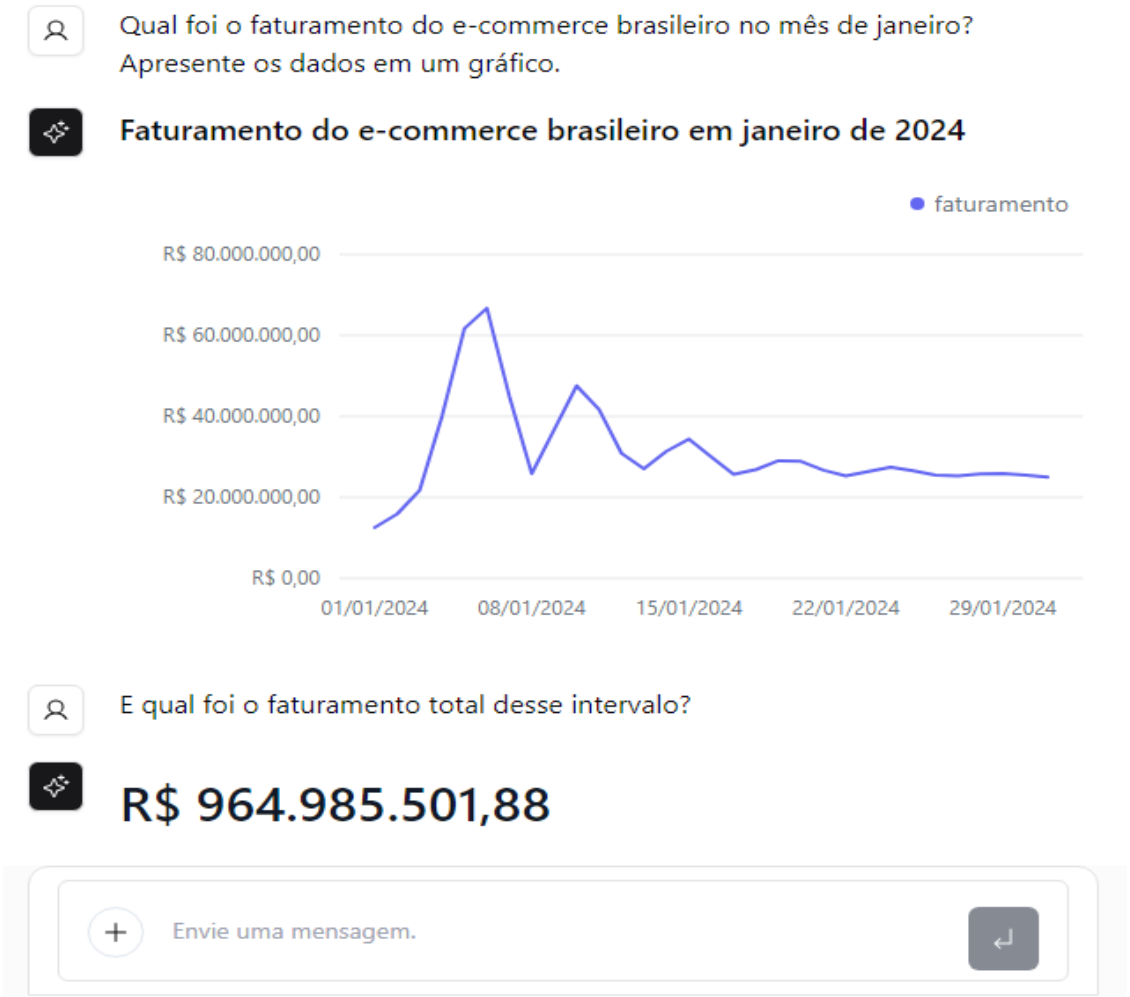


Fonte: Produzido pelo autor.

Além disso, o chatbot é capaz de responder em diversos formatos específicos. Isso inclui a geração de tabelas para organizar dados de forma clara e estruturada, a apresentação de números destacados (big numbers) para enfatizar métricas importantes, e a criação de gráficos para visualizar informações de maneira intuitiva e

compreensível. Esses componentes personalizáveis permitem que o chatbot atenda a uma ampla gama de necessidades de apresentação de informações, tornando a interação mais intuitiva e relevante. Por exemplo, ao solicitar uma análise de vendas, o usuário pode receber um gráfico de tendências, uma tabela detalhada de vendas por região ou um número destacado mostrando o total de vendas do mês. Esses recursos visuais são ilustrados na Figura 4.

Figura 4 – Interface do chatbot com resultados no formato de componentes customizáveis



Fonte: Produzido pelo autor.

Para validar os resultados obtidos pelos modelos, foi utilizado uma abordagem que envolveu o uso de uma planilha contendo um conjunto de perguntas e suas respostas esperadas. Essa planilha foi essencial para comparar as respostas geradas pelo modelo com as respostas corretas, concebida com base no comportamento da

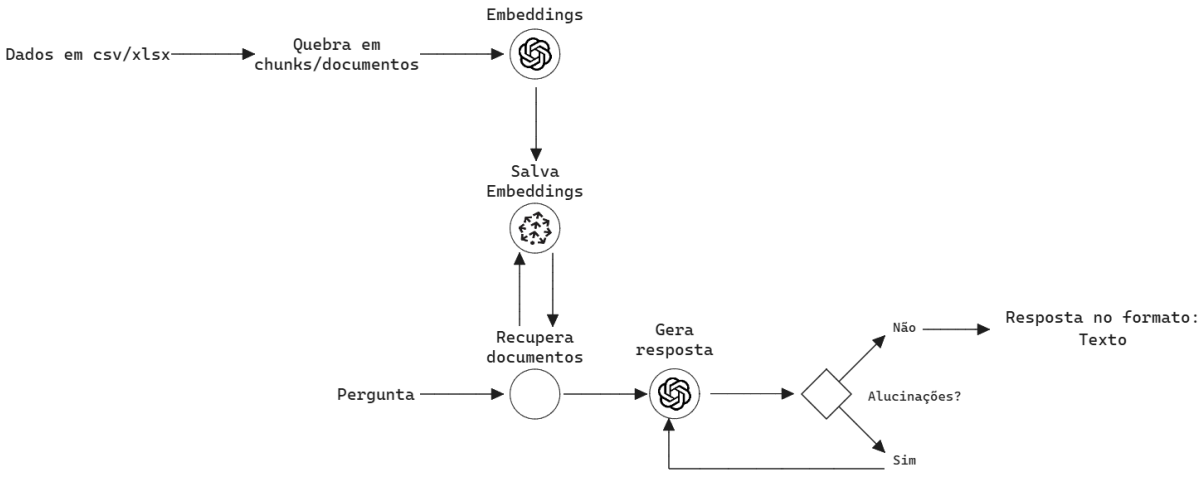
ferramenta TruthfulQA, que é um benchmark composto por um conjunto de perguntas elaboradas especificamente para induzir respostas enganosas, representando um desafio significativo para a avaliação da precisão e confiabilidade dos modelos de IA (LIN; HILTON; EVANS, 2021).

3. RESULTADOS E DISCUSSÃO

Primeiramente, é importante destacar que a análise do comportamento do agente utilizando RAG juntamente com técnicas como a Self-RAG foi conduzida com o objetivo de avaliar sua eficácia em lidar com diferentes tipos de perguntas. Em cenários que envolviam perguntas simples e diretas, o agente demonstrou um desempenho excepcional, gerando respostas precisas e contextualmente adequadas. No entanto, em situações mais complexas, como consultas que requerem relacionamentos complexos ou processamento de longas sequências temporais, o agente apresentou algumas limitações.

Um dos principais desafios identificados foi a falta de algumas informações relevantes recuperadas do contexto, o que limitava a capacidade do modelo de gerar respostas precisas e contextualmente adequadas. Esse problema resultava na geração de respostas com alucinações, variando em cada ciclo de interação. Esta deficiência é evidenciada na análise apresentada na Figura 5.

Figura 5 – Modelo utilizando RAG e Self-RAG

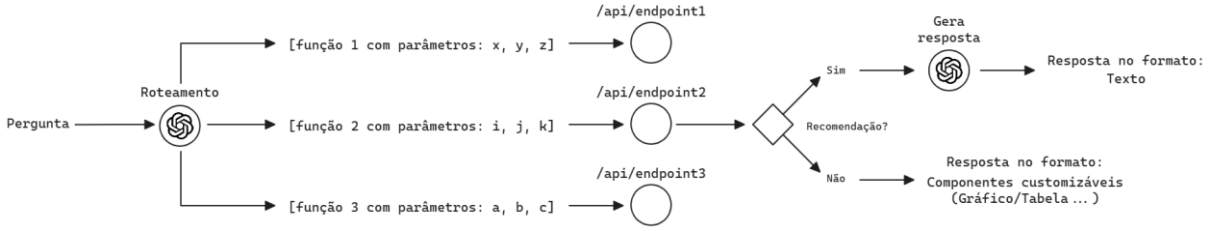


Fonte: Produzido pelo autor.

Em segundo lugar, a estratégia de utilizar ferramentas (tools) e uma abordagem de seleção similar ao Adaptative-RAG Jeong *et al.* (2024) foi implementada, consistindo na decomposição da pergunta em seus componentes básicos para a obtenção de parâmetros e, por meio de um roteamento, identificou-se qual rota se encaixava com os parâmetros estabelecidos. Dessa forma, as requisições eram direcionadas de acordo com a necessidade específica do usuário, permitindo a obtenção de dados relevantes para o retorno da resposta ao usuário.

As respostas geradas foram além das recomendações que o sistema era capaz de oferecer ao usuário com base nos dados obtidos e na análise do contexto da interação. Além disso, o sistema permitiu a entrega de componentes customizáveis, como gráficos e tabelas, que facilitaram a visualização e a compreensão das informações. A Figura 6 ilustra o funcionamento desse processo, evidenciando a integração entre o roteamento inteligente e as APIs utilizadas para a obtenção e o tratamento dos dados.

Figura 6 – Modelo utilizando tools e rotas adaptativas



Fonte: Produzido pelo autor.

A análise dos resultados revelou um desempenho insatisfatório do agente ao utilizar o modelo RAG com a técnica Self-RAG, enfrentando dificuldades em situações com intervalos temporais extensos, muito por causa da falta de informações na hora de recuperar documentos. Essa limitação destaca a complexidade subjacente à compreensão de perguntas que requerem a análise cruzada de dados do e-commerce em intervalos específicos de tempo.

Por outro lado, a estratégia de decompor perguntas e redirecionar requisições de acordo com a necessidade específica do usuário mostrou-se promissora, permitindo a obtenção de dados relevantes e a entrega de respostas mais assertivas e

personalizadas com fácil visualização, demonstrando a eficiência a flexibilidade do sistema de roteamento com integração de APIs.

4. CONSIDERAÇÕES FINAIS

Este projeto evidencia a eficácia da estratégia de decomposição de perguntas e redirecionamento de requisições, resultando em respostas mais precisas e personalizadas para os lojistas. O modelo RAG, aliado a técnicas como o Self-RAG, apresentou limitações ao lidar com perguntas envolvendo intervalos temporais extensos, apontando para a necessidade de melhorias na compreensão contextual de dados temporais. No entanto, a utilização de componentes customizáveis, como gráficos e tabelas, mostrou-se eficiente para a visualização dos dados. Recomenda-se que futuras pesquisas aprimorem a interpretação de informações temporais complexas e continuem a explorar abordagens de roteamento inteligente e integração de APIs, a fim de proporcionar recomendações estratégicas mais relevantes e acessíveis.

5. REFERÊNCIAS

AKYON, Fatih Cagatay. **Paper review 1: ELIZA — A computer program for the study of natural language communication between man and machine**. Medium - NPL CharBolt Survey, 2018. Disponível em: https://medium.com/nlp-chatbot-survey/computational-lingustics-754c16fc7355. Acesso em: 28 nov. 2023.

ASAI, Akari *et al*. Self-rag: Learning to retrieve, generate, and critique through self-reflection. **arXiv preprint arXiv:2310.11511**, 2023. Disponível em: http://arxiv11511.org/abs/2310.. Acesso em: 22 mai. 2024.

BENGFORT, Benjamim, BILBRO, Rebecca; OJEDA, Tony. **Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning.** Sebastopol, CA, USA: O'Reilly Media, 2018.

BIRD, Steven, KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python**. Sebastopol, CA, USA: O'Reilly Media, 2009.

BLETE, Marie-Alice; CAELEN, Olivier. **Developing Apps with GPT-4 and ChatGPT: Build Intelligent Chatbots, Content Generators, and More**. Sebastopol, CA, USA: O'Reilly Media, 2023.

BROWN, Tom *et al*. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877-1901, 2020. Disponível em: http://arxiv.org/abs/2005.14165. Acesso em: 21 mai. 2024.

CHRISTIANO, Paul F. *et al*. Deep reinforcement learning from human preferences. **Advances in neural information processing systems**, v. 30, 2017. Disponível em: https://arxiv.org/abs/1706.03741. Acesso em: 28 nov. 2023.

DAMERAU, Fred J.; INDURKHYA, Nitin (Orgs.). **Handbook of natural language processing**. 2. ed. Filadélfia, PA, USA: Chapman & Hall/CRC, 2010.

DENG, Li; LIU, Yang. **Deep learning in natural language processing**. Springer, 2018.

DEVLIN, Jacob *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Disponível em: http://arxiv.org/abs/1810.04805. Acesso em: 20 mai. 2024.

E-COMMERCE BRASIL. **Com pandemia, e-commerce mais que dobra e já chega a 21% das vendas**. InfoMoney, 2021. Disponível em: https://www.ecommercebrasil.com.br/noticias/com-pandemia-e-commerce-mais-que-dobra-e-ja-chega-a-21-das-vendas. Acesso em: 28 nov. 2023.

FACELI, Katti; LORENA, Ana C.; GAMA, João; *et al*. **Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina**. São Paulo: Grupo GEN, 2021. *E-book*. ISBN 9788521637509. Disponível em:

https://integrada.minhabiblioteca.com.br/#/books/9788521637509. Acesso em: 25 nov. 2023.

GÉRON, Aurélien. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes**. Rio de Janeiro: Editora Alta Books, 2021. *E-book*. ISBN 9786555208146. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786555208146. Acesso em: 28 nov. 2023.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. Cambridge: MIT Press, 2016. Disponível em: http://www.deeplearningbook.org. Acesso em: 19 mai. 2024.

GOOGLE. **Documentação do Dialogflow Vertex: Agents**. Disponível em: https://cloud.google.com/dialogflow/vertex/docs/concept/agents. Acesso em: 22 maio 2024.

GUU, Kelvin *et al*. Retrieval augmented language model pre-training. In: **International conference on machine learning**. PMLR, 2020. p. 3929-3938. Disponível em: http://arxiv.org/abs/2002.08909. Acesso em: 22 mai. 2024.

JEONG, Soyeong *et al*. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. **arXiv preprint arXiv:2403.14403**, 2024. Disponível em: http://arxiv.org/abs/2403.14403. Acesso em: 27 mai. 2024.

KARPATHY, Andrej. **The Unreasonable Effectiveness of Recurrent Neural Networks**. Blog, 2015. Disponível em: http://karpathy.github.io/2015/05/21/rnn-effectiveness. Acesso em: 27 nov. 2023.

KUBLIK, Sandra; SABOO Shubham. **GPT-3: Building Innovative NLP Products Using Large Language Models**. Sebastopol, CA, USA: O'Reilly Media, 2022.

LEWIS, Patrick *et al*. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in Neural Information Processing Systems**, v. 33, p. 9459-9474, 2020. Disponível em: http://arxiv.org/abs/2005.11401. Acesso em: 22 mai. 2024.

LIN, Stephanie; HILTON, Jacob; EVANS, Owain. Truthfulqa: Measuring how models mimic human falsehoods. **arXiv preprint arXiv:2109.07958**, 2021. Disponível em: https://arxiv.org/abs/2109.07958. Acesso em: 29 nov. 2023.

LIU, Xiaoxia *et al*. Prompting frameworks for large language models: A survey. **arXiv preprint arXiv:2311.12785**, 2023. Disponível em: https://arxiv.org/abs/2311.12785. Acesso em: 27 nov. 2023.

MARTINS, Júlio S.; LENZ, Maikon L.; SILVA, Michel Bernardo Fernandes Da; *et al*. **Processamentos de Linguagem Natural**. Porto Alegre: Grupo A, 2020. *E-book*. ISBN 9786556900575. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786556900575. Acesso em: 27 nov. 2023.

MATIAS-PEREIRA, José. **Manual de Metodologia da Pesquisa Científica**. São Paulo: Grupo GEN, 2016. *E-book*. ISBN 9788597008821. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788597008821. Acesso em: 29 nov. 2023.

MAYNEZ, Joshua *et al*. On faithfulness and factuality in abstractive summarization. **arXiv preprint arXiv:2005.00661**, 2020. Disponível em: http://arxiv.org/abs/2005.00661. Acesso em: 22 mai. 2024.

NORVIG, Peter. **Inteligência Artificial**. São Paulo: Grupo GEN, 2013. *E-book*. ISBN 9788595156104. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595156104. Acesso em: 28 nov. 2023.

OPENAI. **Fine-Tuning Guide**. Disponível em: https://platform.openai.com/docs/guides/fine-tuning. Acesso em: 24 nov. 2023.

OPENAI *et al*. GPT-4 Technical Report. **arXiv preprint arXiv:2303.08774**, 2023. Disponível em: http://arxiv.org/abs/2303.08774. Acesso em: 22 mai. 2024.

PINECONE. **Pinecone**. Disponível em: https://www.pinecone.io/. Acesso em: 27 maio 2024.

SAS. **Brasil é o país mais avançado da América Latina no uso de inteligência artificial**. SAS Institute Inc., 2022. Disponível em: https://www.sas.com/pt_br/news/press-releases/2022/october/brasil-e-o-pais-mais-avancado.html. Acesso em: 24 nov. 2023.

SCHICK, Timo *et al*. Toolformer: Language models can teach themselves to use tools. **Advances in Neural Information Processing Systems**, v. 36, 2023. Disponível em: https://arxiv.org/abs/2302.04761. Acesso em: 22 mai. 2024.

SCHLICHT, Matt. **The Complete Beginner's Guide To Chatbots.** Chatbots Magazine, 2016. Disponível em: https://chatbotsmagazine.com/the-complete-beginner-s-guide-to-chatbots-8280b7b906ca. Acesso em: 26 nov. 2023.

TEAM, Gemini *et al*. Gemini: a family of highly capable multimodal models. **arXiv preprint arXiv:2312.11805**, 2023. Disponível em: http://arxiv.org/abs/2312.11805. Acesso em: 22 mai. 2024.

TOUVRON, Hugo *et al*. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023. Disponível em: http://arxiv.org/abs/2307.09288. Acesso em: 29 nov. 2023.

TURING, Alan M. Computing Machinery and Intelligence. **Mind**, v. LIX, p. 433-460, 1950. Disponível em: https://redirect.cs.umbc.edu/courses/471/papers/turing.pdf. Acesso em: 29 nov. 2023.

VASWANI, Ashish *et al*. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017. Disponível em: http://arxiv.org/abs/1706.03762. Acesso em: 20 mai. 2024.

VERCEL. **Next.js**. Disponível em: https://nextjs.org/. Acesso em: 27 mai. 2024.

VERCEL. **Vercel SDK Documentation: Introduction**. Disponível em: https://sdk.vercel.ai/docs/introduction. Acesso em: 27 mai. 2024.

WAZLAWICK, Raul S. **Metodologia de Pesquisa para Ciência da Computação**. São Paulo: Grupo GEN, 2020. *E-book.* ISBN 9788595157712. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788595157712. Acesso em: 29 nov. 2023.

WEIZENBAUM, Joseph. **ELIZA—a computer program for the study of natural language communication between man and machine**. Association Computing Machinery. Copyright ACM 1966. Disponível em: https://dl.acm.org/doi/10.1145/365153.365168. Acesso em: 26 nov. 2023.

YOU, Weiqiu; SUN, Simeng; IYYER, Mohit. Hard-coded Gaussian attention for neural machine translation. **arXiv preprint arXiv:2005.00742**, 2020. Disponível em: https://arxiv.org/abs/2005.00742. Acesso em: 20 mai. 2024.

ZHANG, Aston; LIPTON, Zachary C.; LI, Mu; SMOLA, Alexander J. **Dive into Deep Learning**. Cambridge: Cambridge University Press, 2023. Disponível em: https://D2L.ai. Acesso em: 19 mai. 2024.