INSTITUTO FEDERAL DE CIÊNCIA E TECNOLOGIA DE PERNAMBUCO
Campus Paulista
Análise e Desenvolvimento de Sistemas

LUIZ FELIPE DE ANDRADE RODRIGUES

GYOVANA LUCENA BONINI

Desenvolvimento de um Bot Inteligente para Atendimento ao Cliente com Base em Perguntas Frequentes Usando o Modelo GPT com Fine-Tuning

Paulista

2025

LUIZ FELIPE DE ANDRADE RODRIGUES

GYOVANA LUCENA BONINI

Desenvolvimento de um Bot Inteligente para Atendimento ao Cliente com Base em Perguntas Frequentes Usando o Modelo GPT com Fine-Tuning

Monografia apresentada ao curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco – Campus Paulista, como requisito parcial para obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas. Orientador: Prof. Dr. Rodrigo Cesar Lira da Silva

Paulista

2025

R696d Rodrigues, Luiz Felipe de Andrade.

2025 Desenvolvimento de um Bot inteligente para atendimento ao cliente com base em perguntas frequentes usando o modelo GPT com Fine-Tuning / Luiz Felipe de Andrade Rodrigues; Gyovana Lucena Bonini. – Paulista, PE: Os Autores, 2025.

42f. il. Color.

TCC (Curso Superior Tecnológico em Análise e Desenvolvimento de Sistemas) – Instituto Federal de Pernambuco, 2025.

Inclui Referências

Orientador: Professor Drº Rodrigo Cesar Lira da Silva

1. ChatBot. 2. FAQ. 3. ChatGPT 4. CRM. 5. Fine-Tuning. I. Título. II. Silva, Rodrigo Cesar Lira da (orientador). III. Instituto Federal de Pernambuco.

CDD 006.3

Catalogação na fonte: Bibliotecário Cristian do Nascimento Botelho CRB4/1866

AGRADECIMENTOS

Agradecemos primeiramente a Deus, por nos conceder saúde, força e sabedoria ao longo dessa caminhada.

Às nossas famílias, pelo apoio incondicional, incentivo e compreensão em todos os momentos.

À minha mãe, Marcelle Sales, que infelizmente não está mais entre nós, mas cuja força, amor e ensinamentos continuam presentes em cada passo que dou. Este trabalho é também por ela, que sempre acreditou em mim.

Aos colegas e amigos, pelo companheirismo e troca de experiências durante toda a graduação.

Ao nosso orientador, Rodrigo Lira, pela paciência, dedicação e orientação técnica que foram fundamentais para o desenvolvimento deste trabalho.

Por fim, agradecemos a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

RESUMO

É notável a presença cada vez maior das evoluções tecnológicas nas mais diversas áreas de estudos, produção e prestação de serviços presentes na sociedade, dentre elas se destaca a área de atendimento ao cliente. Tendo em vista a necessidade de otimizar a jornada dos clientes e aprimorar a fidelização e satisfação dos mesmos, bem como ficar a par com a concorrência, cada vez mais empresas investem em tecnologias voltadas ao atendimento, dentre elas se destacam os *chatbots*, os quais permitem a automação de certas etapas da jornada do cliente ao mesmo tempo que buscam manter a qualidade e satisfação almejadas. Este trabalho aborda o desenvolvimento de um bot inteligente para atendimento, com base em técnicas de *fine-tuning* aplicadas sobre o modelo GPT. A base de treinamento é preenchida com perguntas frequentes (FAQs), extraídas de interações reais de clientes de um sistema de Gestão de Relacionamento com o Cliente (CRM). A proposta visa atender a demanda do sistema de oferecer uma experiência satisfatória e eficiente no atendimento, se adequando assim aos aspectos necessários para manter a competitividade no mercado de atendimento ao cliente.

Palavras-chave: ChatBot, FAQ, ChatGPT, CRM, *fine-tuning*.

ABSTRACT

The growing presence of technological advancements across various fields of study, production, and service delivery in society is remarkable, and among these, the customer service sector stands out. Given the need to optimize the customer journey and enhance customer loyalty and satisfaction while also keeping up with its competition, more and more companies are investing in technologies aimed at improving customer service. Among these technologies, chatbots are particularly noteworthy, as they allow the automation of certain stages of the customer journey while striving to maintain the desired quality and satisfaction. This work addresses the development of an intelligent customer service bot, based on *fine-tuning* techniques applied on the GPT model. The training dataset consists of frequently asked questions (FAQs), extracted from real customer interactions within a Customer Relationship Management (CRM) system. The proposal aims to meet the system's demand to deliver a satisfactory and efficient service experience, thereby aligning with the necessary aspects to remain competitive in the customer service market.

Key-words: ChatBot, FAQ, ChatGPT, CRM, *fine-tuning*.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

IA - Inteligência Artificial
CRM - Customer Relationship Management
GPT - Generative Pre-trained Transformer
PEFT - Parameter-Efficient Fine-Tuning
API - Application Programming Interface
CEO - Chief Executive Officer
AGI - Artificial General Intelligence
NPL - Natural Language Processing
PLM - Pre-trained Language Models
LLM - Large Language Model
FAQ - Frequently Asked Questions
XSLX - Microsoft Excel Spreadsheet
JSON - JavaScript Object Notation
JSONL - JSON Lines

SUMÁRIO

1. INTRODUÇÃO

Um dos pontos de vista do marketing de relacionamento, conceito derivado do marketing tradicional, trata-se da adequação da metodologia de atendimento e comunicação de modo a introduzir a tecnologia em seus processos como uma vantagem competitiva, possibilitando o aprimoramento da colaboração e captação de valores entre cliente e empresa (Gordon, 1999).

O conceito de adaptação a novos modelos de acompanhamento também é retratado como um aspecto advindo da globalização, que ocasiona uma adaptação das organizações para manter a competitividade no mercado (Friedman, 2005). A necessidade de fidelizar e satisfazer os consumidores leva as empresas a investirem em canais de comunicação acessíveis, que atendam à diversidade do público e sejam centrados na experiência do cliente, pois as redes sociais são consideradas um caminho promissor para aplicação de estratégias de comunicação para com clientes, considerando que os mesmos constantemente buscam informações sobre as empresas através de meios virtuais (Torres, 2009).

Para ilustrar o potencial dessas plataformas, pode-se citar o WhatsApp, que foi fundado em 2009 e comprado pela META em 2014. Atualmente, o WhatsApp possui uma base de aproximadamente 120 milhões de usuários no Brasil, estando presente em aproximadamente 99% dos aparelhos no país, conforme pesquisa da Panorama Mobile Time/Opinion Box.

Ainda de acordo com o site, foi destacada a ferramenta Whatsapp Business, criada pela META devido a visualização da necessidade de adaptar a plataforma para um cenário que já apresentava sinais de crescimento: a comunicação entre empresas e clientes. O crescimento deste contexto trouxe a apresentação de novas necessidades de aprimoramento para a qualidade e eficácia do atendimento, não só levando em conta o cenário de venda inicial, mas toda a jornada de fidelização e ciclo de vida do cliente, seja este cenário ambientado em empresas de grande ou pequeno porte.

Levando em conta este panorama, empresas e organizações têm investido em inovações tecnológicas, como inteligências artificiais (IA), utilizando-a como um suporte para tornar o atendimento ágil. Sobre esse tema, Auana Mattar destaca:

> “A integração de Inteligência Artificial nas operações de atendimento, movimento conhecido como Contact Center AI, tende a evoluir ainda mais nos próximos anos. Segundo dados do Google Cloud, houve um aumento de 48%, entre 2020 e de 2022, nas chamadas de suporte ao cliente via contact centers”.

Ferramentas como *Chatbots*, definidas como programas computacionais destinados a promover conversas com humanos, bem como agentes de conversação artificial, robôs inteligentes e assistentes digitais (Adamopoulou; Moussiades, 2020) são cada vez mais utilizadas para o suporte ao cliente estão em constante expansão. Segundo dados da Juniper Research (2023), os gastos mundiais do varejo com *Chatbots* devem subir de US$12 bilhões em 2023 para US$72 bilhões até 2028. Isso é um sinal do crescente uso de *Chatbots* em escala global.

O ChatGPT, um recurso da OpenAI, é o que mais se destaca, alcançando a marca de 1 milhão de usuários cinco dias após seu lançamento em novembro de 2022 (Demandsage, 2024). Atualmente, o ChatGPT tem mais de 200 milhões de usuários ativos por semana, de acordo com a demanda de 2024 (Demandsage, 2024).

1.1. JUSTIFICATIVA

Visto o cenário da crescente presença de avanços tecnológicos nos mecanismos de atendimento ao cliente, cuja adoção promove consideráveis vantagens tanto em âmbitos de concorrência no mercado (Qian; Wang, 2017) quanto na satisfação dos clientes (Patil; Digital; India, 2025), uma das ferramentas de destaque são os *Chatbots*, cada vez mais adotados pelas empresas em suas estratégias (Nicolescu; Tudorache, 2022). Ainda abrangendo este cenário, é destacável também a presença da aplicação de técnicas de inteligência artificial voltadas ao aprimoramento dos atuais mecanismos voltados ao atendimento.

Tendo em vista a conjuntura apresentada, unindo a demanda interna da Omnize de aperfeiçoar e deixar mais atraente seu sistema de Gestão de Relacionamento com o Cliente (*Customer Relationship Management* – CRM), especialmente no que diz respeito à oferta de um *chatbot* capaz de fornecer respostas dentro do fluxo de atendimento ao cliente, estabeleceu-se propósito deste estudo. Assim, buscou-se investigar e implementar as técnicas apresentadas, de modo a fortalecer a competitividade da empresa e aprimorar a satisfação dos clientes atendidos.

1.2. OBJETIVO

Este trabalho tem como objetivo o desenvolvimento de um *bot* inteligente para atendimento ao cliente, utilizando perguntas frequentes dos clientes ativos do sistema como base para criação de um *fine-tuning* (Ajuste fino) no modelo GPT 3.5.

1.3. ESTRUTURA DO DOCUMENTO

Este trabalho é composto por seis capítulos, cada um abordando aspectos essenciais para o desenvolvimento do bot usando o *fine-tuning* do ChatGPT.

O primeiro capítulo contextualiza o cenário atual sobre o envolvimento de técnicas de atendimento ao cliente e evoluções tecnológicas, destacando a importância e os benefícios que a união de ambos podem trazer às empresas.

O segundo capítulo dá continuidade a esse raciocínio, especificando os temas relacionados ao atendimento e aprofundando a análise sobre o uso de técnicas de inteligência artificial. São apresentados conceitos e aplicações de Aprendizagem Profunda e Redes Neurais e, por fim, o ChatGPT é introduzido com a apresentação de trabalhos relacionados à temática.

O terceiro capítulo discorre sobre a metodologia idealizada para a realização do projeto, envolvendo o formato e construção da base de dados, a escolha do modelo base e a execução do treinamento e das validações.

O quarto capítulo apresenta as métricas selecionadas para qualificação do modelo e os resultados obtidos após a avaliação, detalhando os parâmetros considerados e os valores resultantes.

O quinto capítulo dedica-se à análise e reflexão sobre os resultados obtidos, bem como aos pontos positivos e negativos observados ao longo do processo, destacando possíveis caminhos alternativos e decisões que poderiam aprimorar ou facilitar o alcance do objetivo inicial.

Por fim, o sexto capítulo reúne as referências utilizadas para embasamento teórico e direcionamento na execução do trabalho.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados os conceitos primários utilizados neste trabalho. Esses conceitos são importantes para a compreensão básica do que será desenvolvido.

2.1. ATENDIMENTO AO CLIENTE

A construção de uma boa relação com clientes, desde a apresentação do produto ou serviço até o processo de acompanhamento pós-venda, é de grande importância para que a empresa mantenha boa reputação e potencialize a garantia do crescimento de seus negócios.

Segundo Rodrigo Bocchi, CEO da Delfia, em sua contribuição no prefácio do livro *Jornada da experiência do cliente*, ele destaca que é importante não apenas tomar medidas pontuais para aprimorar o atendimento, mas manter constante a qualidade em todas as etapas existentes na trajetória do cliente (Bocchi, 2022).

Ainda segundo Bocchi, os clientes querem ser apoiados no processo decisório de compra com base em conhecimento e melhores tecnologias, retomando a importância do desenvolvimento e aplicação de inovações como grandes aliados a construção de relações sólidas com bons clientes e a garantia da satisfação dos mesmos (Bocchi, 2022). Dentro do escopo de inovações tecnológicas voltadas especialmente à área de atendimento, é imprescindível citar o crescimento da Inteligência Artificial e de suas técnicas aplicadas a este setor.

Além do favorecimento da construção de uma boa relação com o cliente, estudos apontam que tecnologias baseadas em Inteligência Artificial aplicadas no mercado contribuem significativamente para a otimização dos processos de venda, seja no pré-venda, durante ou no pós-venda, como é demonstrado, por exemplo, em uma solução aplicada na estratégia de vendas do Linkedin cujo objetivo era focado na renovação de contas vendidas para empresas, onde registrou um aumento de cerca de 8,08% em sua retenção ao utilizar um mecanismo de priorização de contas baseado em IA (Jena; Yang; Tan, 2023).

Ainda em relação à colaboração da IA em vendas, de acordo com um relatório da McKinsey & Company (2021), empresas que adotaram IA no atendimento

conseguiram aumentar a satisfação do cliente em até 30% e reduzir custos operacionais em 20%. Levando em conta estes dados, são notáveis os diversos benefícios da inclusão de técnicas de IA nas estratégias de atendimento, motivando a maior adesão a este meio por parte de empresas e negócios de todos os portes e áreas.

2.3. INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é um campo da computação que tem como objetivo o desenvolvimento sistemas capazes de simular situações e realizar atividades que normalmente exigiriam inteligência humana, tais como raciocínio, aprendizado e tomada de decisões simples ou complexas, como cita o pesquisador João Teixeira, em IA fala-se da elaboração de programas de computador que são ‘modelos’ da capacidade humana de raciocinar, de enxergar, de falar, etc. O conceito de Inteligência Artificial, desenvolvido na década de 40 por Walter Pitts e Warren McCulloch, permanece até os dias atuais, tendo sido amplamente aprimorado e ramificado em várias técnicas, incluindo Aprendizagem de Máquina, Processamento de Linguagem Natural (NLP), Visão Computacional, Robótica e Computação Cognitiva. Essas técnicas são empregadas em várias áreas com objetivos variados, que vão desde a execução de tarefas simples até simulações complexas, exigindo grande capacidade de processamento.

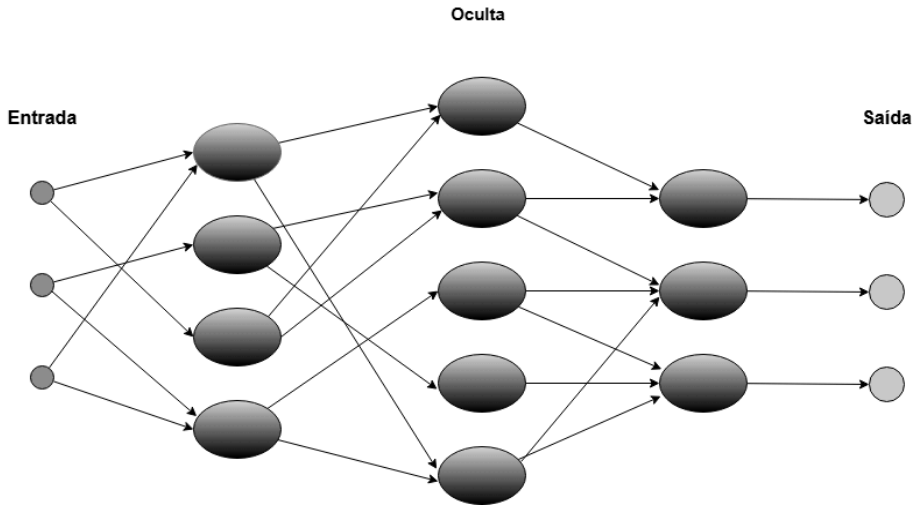
2.3.1. REDES NEURAIS ARTIFICIAIS

Redes Neurais Artificiais são algoritmos de aprendizagem de máquina projetados para se aproximarem às decisões tomadas pelo cérebro humano. Existem diversas técnicas classificadas como Redes Neurais, dentre elas Perceptron, Perceptron Multicamadas (MLP), Redes Neurais Profundas, Redes Neurais Convolucionais (CNN), Redes Neurais Recorrentes (RNN), Transformadores (Transformers) e Redes Neurais Generativas.

As redes neurais simulam a síntese de informações realizada pelo cérebro humano para aprender padrões, processando dados por meio de neurônios artificiais formados por camadas de nós, conforme ilustrado na figura 1. Essas camadas são divididas em: camada de entrada, responsável por tratar dados de vários formatos,

assim como imagens, textos e números. As camadas ocultas, responsáveis por processar dados brutos de entrada em cada neurônio calculando e ajustando os pesos na busca de padrões; camada de saída, responsável pelo resultado do processamento dos dados.

Figura 1: Representação dos nós e fluxo dos dados introduzidos na entrada, passando pelas camadas ocultas e chegando a saída.



Fonte: Elaborada pelos autores.

Após o resultado, ocorre uma avaliação, que é representada como erro. Este erro é determinado ao comparar a saída da camada de saída com o resultado previsto. O *backpropagation*, método crucial para treinar tanto modelos de redes neurais simples quanto as complexas arquiteturas de redes neurais profundas, é utilizado para ajustar o modelo. Ele faz o cálculo ao comparar a saída da rede com o valor previsto. Após isso, o *backpropagation* propaga o erro camada por camada, de trás para frente, estabelecendo os pesos das conexões que precisam ser modificados (Shihab, 2006).

Esse ajuste é realizado pelo algoritmo de otimização gradiente descendente, que tem como função reduzir os erros entre os resultados previstos e os reais de maneira eficaz. Portanto, a cada ciclo os parâmetros são otimizados, melhorando sua capacidade de predição, podendo assim ser adequados para uso em várias áreas.

2.3.2 MODELOS DE LINGUAGEM DE GRANDE ESCALA (LLMs)

Large Language Model (LLM) trata-se de uma evolução de modelos de linguagem pré-treinados ou PLM (*Pre-trained Language Models*) os quais são construídos utilizando técnicas de treinamento de redes neurais do tipo *Transformers* com base em grandes quantidades de dados (Zhao et al. 2023).

Devido a utilização desta técnica, o aumento das capacidades computacionais e a disponibilidade de dados de treinamento, uma grande evolução foi observada no desenvolvimento destes modelos nos últimos tempos, permitindo maior aproximação a capacidades humanas de solução de problemas (Naveed et al. 2024) bem como o aperfeiçoamento de soluções de tarefas no contexto de NPL *(Natural Language Processing)* (Wei et al. 2022).

Dentre os modelos de LLM em destaque, é selecionado para este contexto o modelo GPT (*Generative Pre-trained Transformer*) com foco principal no desenvolvimento e aperfeiçoamento de técnicas de conversação (Zhao et al. 2023), servindo como base para o desenvolvimento do ChatGPT, o qual apresenta excelentes capacidades de comunicação (Zhao et al. 2023).

2.4. CHATGPT

ChatGPT é um modelo de linguagem baseado em inteligência artificial cuja estrutura emprega o modelo GPT, que permite a interação com usuários de forma natural, com foco em interações conversacionais, permitindo a criação de respostas consistentes sobre uma vasta gama de assuntos (Susnjak; 2022).

A OpenAI foi a precursora na criação de modelos de predição baseados em redes neurais, com a criação do primeiro modelo GPT, o GPT-1, em 2018. Desde então, a empresa desenvolveu diversos modelos de GPT, como o GPT-2, em 2019, e o GPT-3, em 2020, que trouxeram melhorias significativas na geração de texto e maior capacidade de compreensão. Por fim, lançou o ChatGPT e seus modelos subsequentes, que foi um marco em sua evolução, representando a capacidade poderosa da combinação das técnicas aplicadas. Com lançamento em 30 de novembro de 2022, o ChatGPT, após 5 dias de sua criação, atingiu a marca de 1 milhão de usuários (Demandsage, 2024).

A seguir, é apresentada uma lista dos principais modelos baseados na arquitetura GPT, acompanhados de seus anos de lançamento.

Modelos de linguagem da OpenAI:

- GPT-1 (2018): Primeiro modelo de linguagem baseado em redes neurais transformadoras. Foi um marco no desenvolvimento de IA generativa.
- GPT-2 (2019): Trouxe avanços significativos na geração de texto coerente, sendo mais robusto e poderoso que seu predecessor.
- GPT-3 (2020): Considerado revolucionário, com 175 bilhões de parâmetros, aumentando drasticamente a capacidade de compreensão e geração de texto.
- GPT-3.5 (2022): Versão aprimorada do GPT-3, utilizada na primeira versão do ChatGPT.
- GPT-4 (2023): Modelo multimodal que combina processamento de texto e imagens, ampliando as capacidades do ChatGPT.

2.4.1. *FINE-TUNING* DO CHATGPT

Fine-tuning é um processo de refinamento de modelos pré-treinados, como GPT-4 e GPT-3.5, entre outros, disponíveis via API (do inglês, *Application Programming Interface*). Esse refinamento é usado para tarefas ou domínios específicos, aproveitando o conhecimento do modelo treinado com uma base diversificada de dados durante o treinamento inicial.

Seu principal propósito é o atendimento ao cliente, interpretação de conteúdo ou pesquisa, a personalização do ChatGPT permite atender às suas necessidades distintas. Esse processo envolve treinar o modelo adicionalmente em um conjunto de dados específico para que ele possa fornecer respostas mais precisas, relevantes ou ajustadas ao contexto desejado.

O custo para usar a API do ChatGPT depende do volume de *tokens*[1] processados. A OpenAI cobra com base no número de *tokens* treinados, enviados (entrada) e gerados (saída). Na própria documentação da OpenAI os preços são catalogados conforme a Tabela 1.

[1] *tokens* são as unidades de texto que o modelo usa para calcular as respostas.

Tabela 1 - Custos definidos pela OpenAI na etapa de *fine-tuning* dos modelos do ChatGPT.

	Entrada	Saída	Treinamento
ChatGPT-3.5	0,008	0,006	0,008
ChatGPT-4.0	0,0003	0,0012	0,003

Preço em dólar (US$) para cada 1000 *tokens*.

Para referência, 1.000 *tokens* correspondem aproximadamente a 750 palavras. O custo total depende do volume de interações e da complexidade dos prompts.

2.5. TRABALHOS RELACIONADOS

Visando compreender e aprofundar a análise da performance de LLMs em cenários de atendimento ao cliente, o estudo de (Ilse; Blackwood, 2024) analisa a viabilidade e eficácia de LLMs com fine-tuning, comparando técnicas de refinamento e estratégias de validação.

Foram avaliados três modelos: GPT-4.0, pela ampla capacidade de adaptação; LLaMA-2, por ter a natureza *open-source*; e Gemini, escolhido por sua arquitetura híbrida. Os autores consideraram diferentes técnicas para treinamento, como o processo tradicional de *fine-tuning*, compreendido por apresentar ótimos resultados ao passo que requer mais recursos computacionais, *Parameter-Efficient Fine-Tuning* (PEFT) o qual otimiza o processo de treinamento e demonstra resultados semelhantes e, por último, o *Domain-Adaptive Pretraining*, técnica que prepara antecipadamente o modelo para a linguagem mais adequada ao seu objetivo.

O Gemini apresentou a melhor performance em acurácia e precisão, embora os demais também tenham tido bons resultados. O estudo concluiu que técnicas de *Domain-Adaptive Pretraining* melhoram o desempenho, mas exigem considerável capacidade computacional.

Seguindo na perspectiva de modelos de larga escala voltados ao atendimento ao cliente, (Tzanis, 2025) apresentou a construção de um chatbot de perguntas e

respostas utilizando a abordagem *Retrieval-Augmented Generation* (RAG), a qual permite reaver informações de fontes externas e integrá-las ao escopo da geração do modelo de forma dinâmica.

Para averiguar a eficácia dos modelos LLM escolhidos, sendo eles llama3-70b, llama3-8b, mixtral-8x7b, gemma-7b-it e gemma2-9b-it, foi utilizado o *RAGAS*[2]. Os resultados obtidos após a aplicação das métricas demonstraram adequação satisfatória das respostas retornadas, porém ao custo de grande capacidade computacional, indicando que, apesar de eficaz, é uma abordagem que apresenta restrições significativas em sua execução.

No contexto de modelos pré-treinados para tarefas de perguntas e respostas, o estudo de (Yiming et al. 2024) apresenta o desenvolvimento de um *chatbot* cujo objetivo é auxiliar na conscientização da população sobre o Papilomavírus Humano (HPV) e a importância da vacinação para prevenção.

A base de perguntas e respostas foi construída com base em artigos científicos selecionados estritamente de acordo com o escopo escolhido, a fim de evitar que informações improcedentes fossem inseridas. Foram escolhidos os modelos GPT-3.5 e GPT-4.0, submetidos sob o mesmo processo de *fine-tuning* e sob a mesma base de perguntas.

A avaliação dos resultados foi realizada também utilizando a abordagem *RAGAS*. O modelo nomeado VaxBot-HPV obteve resultados considerados precisos e relevantes, demonstrando o potencial das LLMs para a execução de tarefas com escopo bem definido.

[2] conjunto de métricas relevantes para avaliação de acurácia e precisão de respostas de modelos que usam *Natural Language Processing*.

3. METODOLOGIA

Neste capítulo será abordada a metodologia para a implementação do *fine-tuning*. Sob esta perspectiva, será abordado a construção da arquitetura, desenvolvimento e avaliação. O objetivo principal deste trabalho é integrar o *fine-tuning* com uma ferramenta de chat.

Nesta etapa foi decidido utilizar o ChatGPT *fine-tuning* por conta da viabilidade técnica, de custo e maior possibilidade de integração com o sistema.

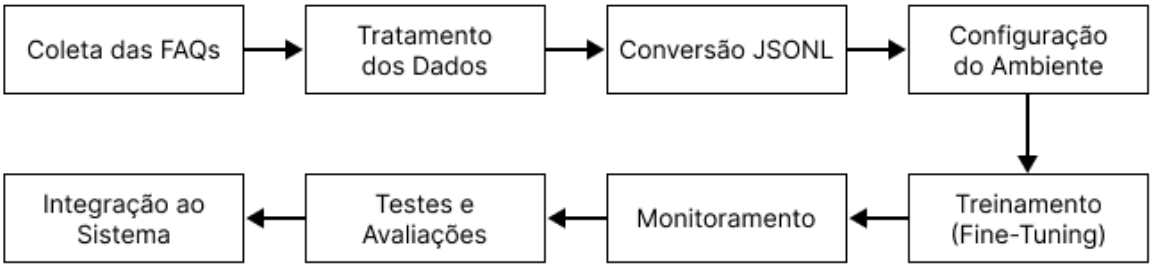
3.1. FLUXOGRAMA DO PROCESSO

Para garantir clareza, a metodologia deste trabalho está estruturada em etapas sequenciais, desde a coleta inicial dos dados até a integração prática do modelo desenvolvido.

A Figura 2 apresenta o fluxo das etapas necessárias para a implementação do projeto de *fine-tuning* do modelo GPT. O processo se inicia com a coleta das perguntas frequentes (FAQs), seguido pelo tratamento e conversão dos dados para o formato JSONL. Em seguida, ocorre a configuração do ambiente e o treinamento do modelo (*fine-tuning*).

Após o treinamento, são realizadas as etapas de monitoramento, testes e avaliação, finalizando com a integração do modelo ajustado ao sistema de atendimento. Esse fluxo visa garantir uma implementação organizada, eficiente e alinhada aos objetivos do projeto.

Figura 2: Representação do fluxo das etapas da implementação do projeto.



Fonte: Elaborada pelos autores.

3.2. FERRAMENTAS E TECNOLOGIAS UTILIZADAS

Para a implementação do *fine-tuning* do ChatGPT, foram utilizadas as seguintes ferramentas e tecnologias:

- **Linguagem de programação:** Python 3.9.
- **Bibliotecas:** *pandas* para manipulação de dados, *json* para conversão de arquivos e *openai* para comunicação com a API da OpenAI.
- **API da OpenAI:** Utilizada para envio dos arquivos JSONL e execução do treinamento do modelo.
- **Ambiente de desenvolvimento:** O código foi implementado no VS Code.
- **Formatos de dados:** Planilhas no formato XLSX, posteriormente convertidas para JSONL.
- **Ferramentas auxiliares:** O Postman foi utilizado para testes de requisições na API.

3.3. PRÉ REQUISITOS PARA UTILIZAÇÃO DA API DA OPENAI

Para possibilitar a utilização dos serviços da OpenAI é necessário o registro de uma conta verificada e possuidora de um método de pagamento ativo e funcional, registrado de acordo com os requisitos da plataforma para possibilitar a realização da cobrança pela geração de *tokens.*

A preparação do modelo, incluindo seu treinamento, teste e utilização, pode ser realizada tanto através da própria interface disponibilizada pela OpenAI quanto através da API disponibilizada pela plataforma, o que viabiliza a integração do modelo com outros sistemas.

Atendendo à necessidade de uma integração completamente configurada através da plataforma, optou-se por possibilitar a configuração do modelo utilizando a API providenciada.

3.4. PREPARAÇÃO DOS DADOS

Nesta seção são descritas as etapas realizadas para a preparação dos dados utilizados no processo de *fine-tuning* do modelo ChatGPT. Isso inclui desde a coleta e

seleção das informações relevantes até o tratamento e formatação dos dados no padrão exigido pela API da OpenAI.

3.4.1. DEFINIÇÃO DO OBJETIVO

A base de treinamento do modelo foi construída considerando perguntas frequentes realizadas por clientes através de canais de Chat vinculados ao CRM, sendo definidas pelo parceiro, o qual utiliza o sistema para a gerência do atendimento ao público. As perguntas selecionadas, assim como o objetivo do modelo, têm um escopo bem definido e atendem a um público consumidor do produto Quickdesk.

As dúvidas e questionamentos foram coletados clientes parceiros da Omnize através da própria plataforma de forma manual, considerando critérios como enquadramento no escopo, frequência apresentada nas interações e coerência na escrita. Junto às perguntas são elaboradas as devidas respostas, as quais podem estar relacionadas a mais de uma pergunta, visando maior acurácia nas etapas de treinamento e validação e no funcionamento do modelo em si.

Esse processo tem o objetivo de estreitar o escopo da base e, ao mesmo tempo, aprimorar a interação com o cliente final, adequando o modelo para oferecimento do melhor suporte perante as necessidades mais predominantes e aumentando a satisfação com relação ao atendimento.

3.4.2. COLETA E SELEÇÃO DOS DADOS

A coleta dos dados utilizados no treinamento do modelo foi realizada por meio de extração de informações disponíveis nos registros de conversas entre clientes e a plataforma de atendimento, acessadas por meio do sistema CRM integrado.

Os dados obtidos foram primeiramente coletados pelo cliente parceiro, o qual alocou um time especializado para inicialmente selecionar e averiguar as informações desejadas e produzir as respostas adequadas. O critério principal escolhido para delimitar as informações e definir o propósito do modelo se baseia nas seguintes premissas:

1. Enquadramento no domínio determinado pelo parceiro;
2. Coerência e objetividade da pergunta;

3. Variedade da expressão de informações (perguntas com sentidos iguais, porém com variação na sintaxe);
4. Remoção de dados sensíveis e conteúdo impróprio;

Após o recolhimento e tratamento das informações, realizados considerando os critérios informados, a base contendo as perguntas e respostas foi enviada pelo parceiro às equipes de suporte e desenvolvimento. Foi então realizada uma nova curadoria, esta com visão voltada ao desenvolvimento.

Foram realizadas alterações voltadas ao aprimoramento da base e a possibilidade de integração mais dinâmica com a plataforma, incluindo valores voltados ao retorno de *tags*[3] de fluxo específicas, as quais funcionam como gatilhos de indicação de caminhos a serem tomados pelo sistema mediante interações e processos internos necessários à continuidade do fluxo de atendimento.

Essas interações foram agrupadas em uma planilha de formato ".XLSX" contendo duas colunas principais: prompt, que representa a pergunta ou solicitação do usuário, e *completion*, que representa a resposta ideal esperada para aquele contexto. Toda a base foi construída a partir de um único tema específico: o processo para se tornar sócio de um clube esportivo. A base resultante contou no total com 500 perguntas, findando a etapa inicial e permitindo a execução dos próximos passos de treinamento e validação.

3.4.3. FORMATAÇÃO DOS DADOS

O tratamento e formatação de dados requeridos para realização da comunicação com a API, considerando o modelo GPT-3.5, é padronizado de acordo com o formato JSONL, o qual se trata de uma extensão do formato JSON. O seu uso se dá por conta do desempenho aprimorado em relação ao JSON comum, considerando a forma de tratamento linha por linha permitida, o que o torna ideal em cenários onde serão trabalhados grandes volumes de dados como no caso de uma

[3] *tags* são sequências de caracteres com propósito identificador, utilizadas para definição ou estreitamento de escopo.

inserção de informações de treinamento e testes de modelos de aprendizagem de máquina.

Após a conversão dos dados para o formato requerido, foi feita a separação dos conjuntos de treinamento e validação de forma à expor o modelo a cenários inéditos em seu processo de aprovação.

A proporção escolhida para a divisão foi: 90% dos dados dedicados ao treinamento e 10% voltados a validação e testes. A escolha desta proporção se deu por conta do tamanho final da base, pois para garantir que o modelo apresentasse um bom desempenho se faz necessária a efetuação do treinamento com a maior quantidade de dados disponível para tal fim.

Inicialmente, os dados separados para o treinamento se encontravam em uma planilha de formato XLSX e, para viabilizar a sua utilização, foi realizada uma conversão utilizando um algoritmo escrito em Python e aplicando as bibliotecas Pandas e JSON. As colunas da planilha, juntamente ao formato da conversão realizada para uso nas requisições estão descritos a seguir:

Colunas da planilha

A planilha utilizada para organizar os dados de treinamento contém duas colunas principais. A primeira, denominada *Prompt*, corresponde à pergunta ou instrução enviada ao modelo, simulando a interação de um usuário. A segunda, *Completion*, representa a resposta esperada para aquele prompt, servindo como referência para o processo de treinamento.

Formato JSONL

Após organizados na planilha, os dados foram convertidos para o formato JSONL, padrão exigido pela API da OpenAI. Cada linha do arquivo contém um objeto com a estrutura messages, composto por uma lista de mensagens. Cada mensagem possui duas propriedades: role, que identifica o papel da mensagem no diálogo (podendo assumir system, user ou assistant), e content, que contém o texto correspondente ao papel definido.

O campo system é utilizado para estabelecer instruções iniciais e delimitar o escopo do modelo; user representa a entrada do usuário; e assistant corresponde à resposta que o modelo deve aprender a reproduzir.

Na figura 3 é apresentado um exemplo do formato final:

Figura 3 – Exemplo de estrutura de mensagens no formato JSON para treinamento do modelo

```
{"role": "system", "content": "Definição de escopo do modelo"},
{"role": "user", "content": "Pergunta do usuário."},
{"role": "assistant", "content":  "Resposta para a pergunta"}
```

Fonte: Elaborada pelos autores.

Para evitar que o modelo treinado produzisse respostas fora do escopo, foram estabelecidas regras diretamente na mensagem do tipo system durante o processo de fine-tuning. No treino, foram utilizadas instruções como: “Você é um assistente especializado em responder perguntas apenas sobre cartões de crédito. Caso a pergunta esteja dentro do escopo, mas não exista um exemplo correspondente na base de treinamento, peça ao usuário que acesse o site oficial para obter mais informações.”

Essa configuração tem a função de restringir o comportamento do modelo, assegurando que perguntas totalmente fora do tema sejam respondidas com mensagens genéricas ou sinalizando explicitamente que o assunto está fora do escopo.

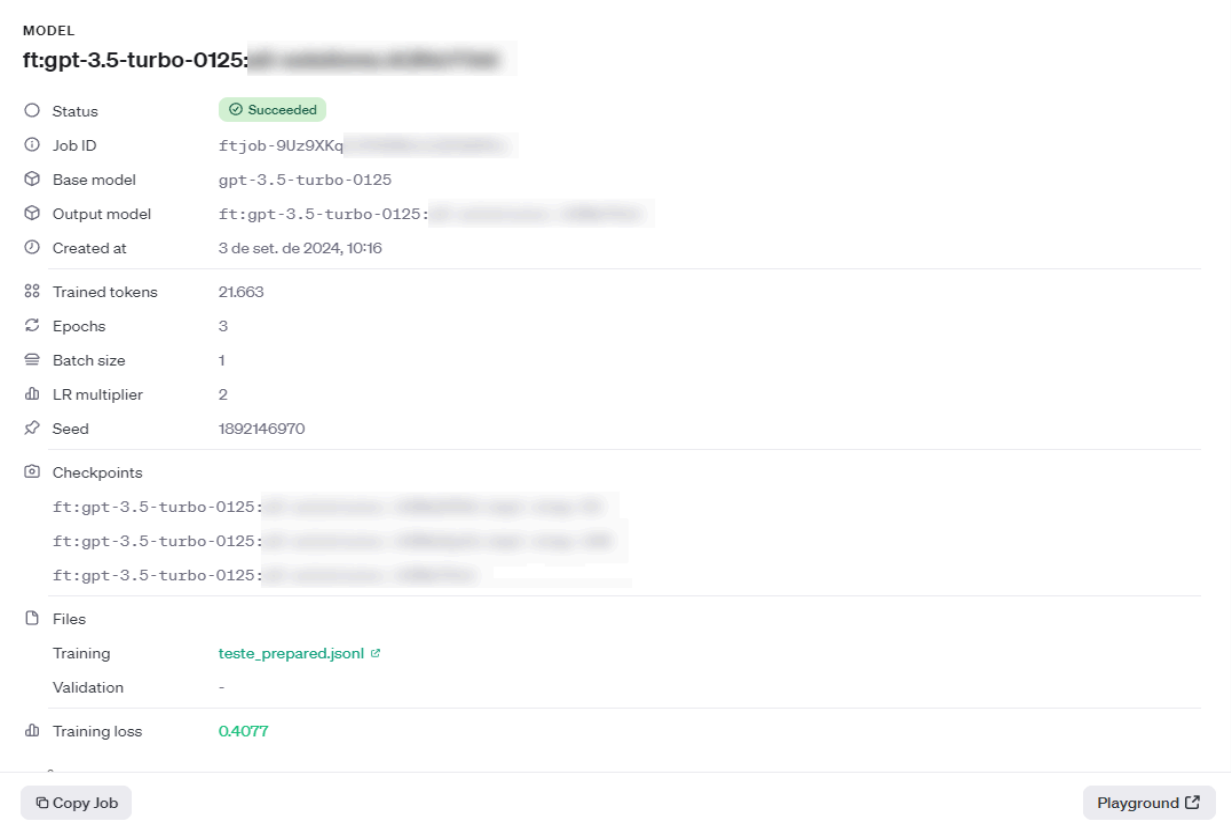
Nos casos raros em que a pergunta pertence ao contexto, mas não possui uma resposta correspondente na base de treinamento, o modelo é instruído a recomendar que o usuário consulte o site oficial para obter informações mais detalhadas.

3.5. CONFIGURAÇÃO DO *FINE-TUNING*

Mediante autenticação via chave de API e utilizando a biblioteca da OpenAI, realiza-se o upload do arquivo no formato JSONL, etapa fundamental na qual são definidos tanto os dados de treinamento quanto o modelo base a ser ajustado, iniciando assim o processo de *fine-tuning*. Após a solicitação para a criação do *fine-tuning* é retornado identificador único, o *‘fine_tune_job_id’*, o qual permite ter acesso ao *status* da criação do *fine-tuning*. Uma vez finalizado com sucesso, é disponibilizado o

nome do modelo ajustado, que segue o padrão '*model=ft:gpt-3.5-turbo-0125:nome_personalizado::ID*', ele permite a utilização do modelo *fine-tuning* em aplicações práticas. A indicação de finalização e sucesso da criação do modelo, juntamente aos parâmetros e informações citadas estão demonstrados na Figura 3.

Figura 4 – Representação da criação bem sucedida do *fine-tuning* na plataforma OpenAI



Fonte: Elaborada pelos autores.

No processo de fine-tuning da OpenAI, os hiperparâmetros de treinamento não são disponibilizados ao usuário. A plataforma realiza todo o ajuste interno de forma automatizada, sem permitir a configuração manual desses parâmetros via API.

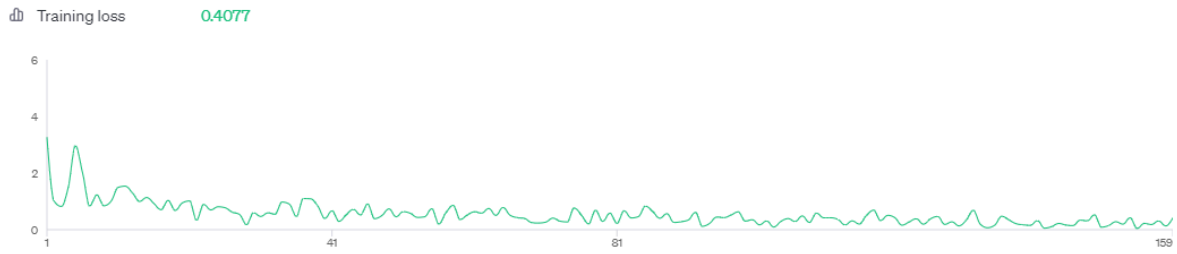
Dessa forma, o treinamento é conduzido utilizando os padrões definidos pela própria OpenAI, que controla internamente aspectos como ciclos de aprendizado, políticas de regularização e estratégias de convergência.

3.6. EXECUÇÃO E MONITORAMENTO DO TREINAMENTO

Após concluir a filtragem e otimização da base de dados, iniciou-se o processo de carregamento e treinamento do modelo. Durante o *fine-tuning*, o modelo é ajustado com base nos dados fornecidos.

Ao final do treinamento, como ilustrado na Figura 4, é gerado um gráfico de curva de *loss* que permite avaliar se o modelo aprendeu corretamente, além de demonstrar a eficácia do processo de ajuste dos pesos durante o treinamento.

Figura 5 – Representação do treino do modelo *fine-tuning* na plataforma OpenAI.



Fonte: Elaborada pelos autores.

3.7. TESTE E INTEGRAÇÃO

Concluído o treinamento, o funcionamento do modelo e sua eficácia foram averiguados por meio de requisições API através do envio de perguntas para a obtenção e avaliação das respostas geradas. Esses testes podem ser realizados de duas formas principais: via Python, utilizando a biblioteca OpenAI, ou por meio de ferramentas de teste de API, como o Postman.

No caso da implementação em Python, a requisição foi realizada por meio do método '*client.chat.completions.create()*', sendo '*client*' uma variável utilizada para receber uma instância da classe *OpenAi* acrescida da definição da propriedade '*api_key*'. Este atributo se refere à chave da API (*API key*), para realizar a cobrança pelo uso e identificar o modelo *fine-tuned*. O parâmetro *messages*, mencionado na seção sobre a formatação JSONL, define a estrutura da conversa com o modelo, incluindo as mensagens com suas respectivas funções (*system, user*) e conteúdos.

Além dele, outros parâmetros são estabelecidos:

- *temperature:* Define se as respostas serão objetivas e previsíveis ou criativas;
- *max_tokens:* Define o máximo de *tokens* de respostas do modelo.

No exemplo demonstrado na Figura 5, o modelo é instruído a se comportar como um assistente especializado em cartões de crédito, respondendo apenas dentro desse escopo:

Figura 6 – Representação do código Python da requisição de teste.

```python
from openai import OpenAI

client = OpenAI(api_key="sk-lv                                                ")

completion = client.chat.completions.create(
  model="ft:gpt-3.5-turbo-0125:                    ",
    messages=[
        {
            "role": "system",
            "content": "Você é um assistente especializado em responder perguntas apenas sobre cartões de crédito."
        },
        {
            "role": "user",
            "content": "Olá"
        }
    ],
    temperature=0.5,
    max_tokens=80
)

print(completion.choices[0].message)
```

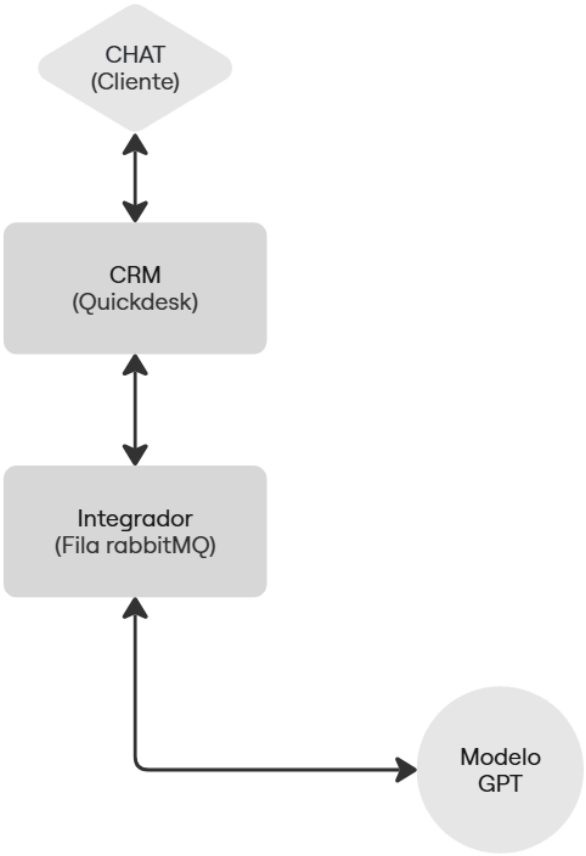
Fonte: Elaborada pelos autores.

3.8. IMPLEMENTAÇÃO

Após a finalização das etapas de treinamento, validação e testes realizados através do consumo da API via código Python, é iniciada a fase de integração com o CRM. A comunicação entre a plataforma e o modelo foi estabelecida por meio de um *driver* responsável por atuar como integrador assíncrono, intermediando o envio de perguntas e o recebimento das respostas do modelo da OpenAI.

Esse integrador é implementado sobre uma fila de mensageria baseada em *RabbitMQ*, o que permite organizar e distribuir as requisições de forma segura e escalável. Assim, cada interação enviada pelo CRM é encaminhada ao modelo GPT e, em seguida, a resposta retornada é reinserida na fila para posterior consumo pelo sistema. Esse fluxo pode ser observado na ilustração apresentada na Figura 6.

Figura 7 – Arquitetura da integração do modelo GPT com o fluxo de automação.



Fonte: Elaborada pelos autores.

Este tipo de integração viabiliza a utilização do novo modelo em um cenário de produção onde interações reais, no contexto de atendimento ao cliente, são realizadas diariamente e as respostas geradas têm como base todos os dados coletados e acrescentados nos arquivos de treinamento do modelo.

4. RESULTADOS

Para avaliar a eficácia do modelo criado, foram consideradas as métricas de validação descritas nas seções seguintes.

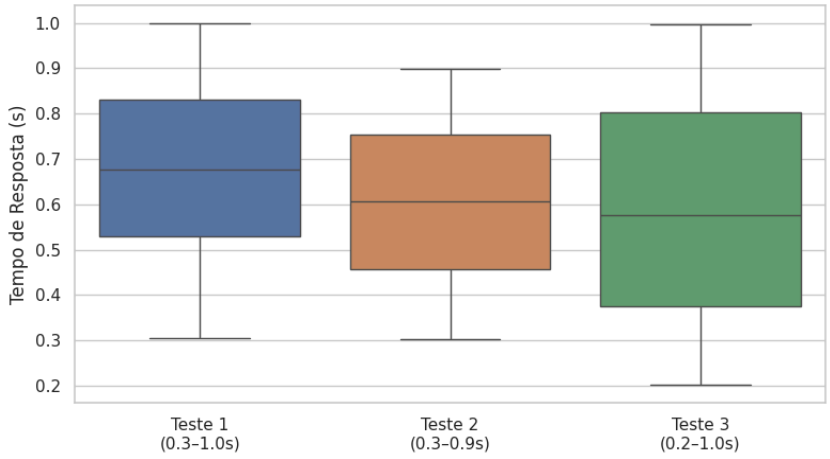
4.1. TEMPO DE RESPOSTA

Nesta etapa foi averiguado o tempo levado para a chegada da resposta após o envio da pergunta através de testes em *scripts* Python. O script utilizado para envio das perguntas, em todos os testes, executou 300 consultas, sendo feito um envio a cada 500 milissegundos e coletando os tempos de resposta apresentados.

Foi avaliada a velocidade de resposta ao longo de três experimentos com os mesmos cenários citados anteriormente. Vale destacar que, por nesta métrica o conteúdo das respostas não ser relevante, também foram incluídas perguntas da base de treinamento, não afetando os resultados finais.

Os intervalos de tempo variaram conforme cada teste: o primeiro entre 0,3s e 1,0s, o segundo entre 0,3s e 0,9s e o terceiro entre 0,2s e 1,0s. A Figura 7 apresenta um gráfico do tipo *boxplot* que ilustra a distribuição dos tempos de resposta em cada uma das simulações (Teste 1, Teste 2 e Teste 3), permitindo observar a mediana, a amplitude e possíveis *outliers*.

Figura 8 – Gráfico de boxplot de comparativo de três testes de tempo de resposta (300 iterações por teste).



Fonte: Elaborada pelos autores.

Mudando de contexto, para mensurar o tempo de resposta considerando a integração completa com o ambiente CRM, foi realizada uma aferição direta durante interações reais na plataforma de chat. A medição foi feita a partir dos registros de log do integrador, que armazenam automaticamente os timestamps de envio da pergunta ao modelo e de recebimento da resposta processada.

O tempo total considerado corresponde ao intervalo entre: o momento em que o CRM encaminha a mensagem para a fila de processamento (timestamp de saída), o momento em que a resposta retornada pelo modelo é efetivamente entregue ao usuário final pelo chat (timestamp de exibição).

Com esse procedimento, observou-se um tempo médio entre 1 s e 4 s por interação, incluindo todo o percurso: CRM → Integrador (RabbitMQ) → API da OpenAI → CRM → Usuário final.

4.2. EQUIDADE E SEMELHANÇA DAS RESPOSTAS

Neste tópico foi avaliada a coerência entre as respostas geradas e as consideradas adequadas a cada pergunta, levando em conta a base inserida.

Vale ressaltar que, para que os resultados sobre este ponto sejam satisfatórios, é interessante que a base de dados esteja bem estruturada e limitada apenas às informações de interesse para a geração de respostas, bem como a boa elaboração e especificidade do *prompt* a ser enviado para a delimitação do objetivo do modelo.

Para realizar a comparação entre as respostas geradas pelo modelo refinado e as consideradas ideais, foi utilizado o conjunto selecionado para validação e então, foi realizado o envio de cada uma das perguntas ao modelo para geração e registro das respostas.

A comparação entre as respostas ideais e as respostas geradas foi feita utilizando a técnica denominada *Cosine Similarity*, a qual funciona através da transformação dos textos em vetores e o cálculo do cosseno do ângulo observado entre os vetores resultantes. A equação utilizada para o cálculo está representada na Figura 8.

Figura 9 – Equação de cálculo implementada pela técnica *Cosine Similarity*.

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Fonte: Kaggle.

Na fórmula apresentada, os valores de *A* e *B* representam os vetores a serem comparados. Inicialmente é calculado o produto entre estes vetores e então, é calculada a divisão do valor resultante pelo módulo ou comprimento do vetor, cujo valor é obtido através da raiz quadrada da soma dos quadrados das coordenadas presentes no vetor. O resultado obtido, considerando este contexto onde todos os valores obtidos são positivos, varia de 0 a 1, sendo 0 a indicação de vetores ortogonais ou sem similaridade e 1 representando vetores exatamente iguais.

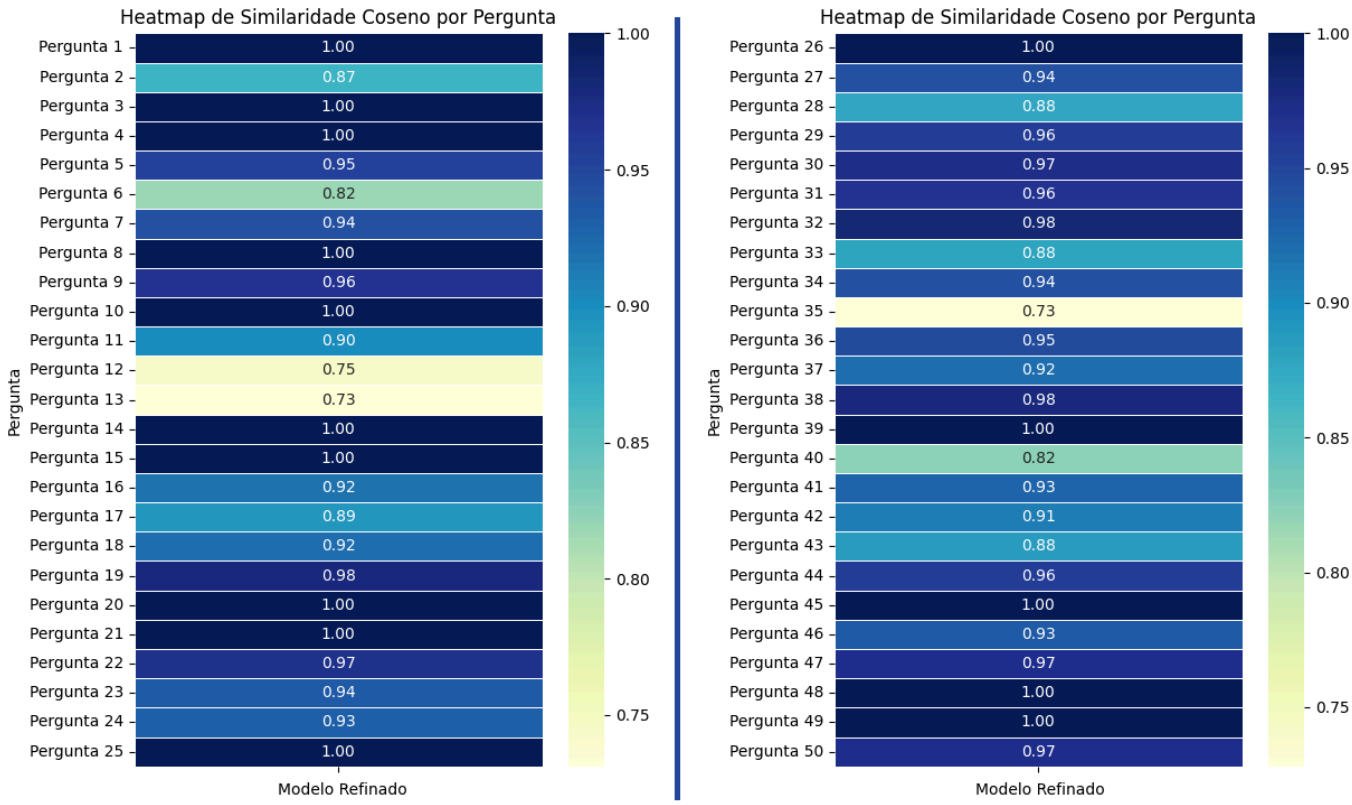
A métrica descrita foi escolhida devido a sua capacidade de identificar semelhanças semânticas abstraindo a magnitude ou tamanho dos vetores, sendo ideal para a identificação correta do contexto da resposta e também permitindo variações em sua geração, incrementando fluidez e dinamismo em conversas com clientes.

Para visualizar de forma clara o desempenho do modelo, foi escolhido o gráfico de mapa de calor, pois o mesmo torna mais perceptível a tendência do modelo num geral, além de destacar possíveis anomalias presentes nos resultados, sendo favorável o seu uso para a quantidade atual de dados de validação e seus resultados.

Na situação atual, o parâmetro ideal definido pela equipe interna de desenvolvimento da empresa executora foi que a quantidade de respostas geradas com similaridade superior a 0.85 deveria ser igual ou maior a 80%, considerando o conjunto de validação.

Os resultados obtidos e apresentados na Figura 9 exibem uma média de similaridade por cosseno aproximada de 0,939 na escala definida e 90% de respostas geradas com similaridade igual ou maior a 0,85. Os valores indicaram boa proximidade entre os resultados esperados e os obtidos ao executar o modelo.

Figura 10 – Gráfico mapa de calor demonstrando os resultados obtidos ao aplicar a similaridade do cosseno, comparando as respostas ideais às geradas pelo modelo refinado.



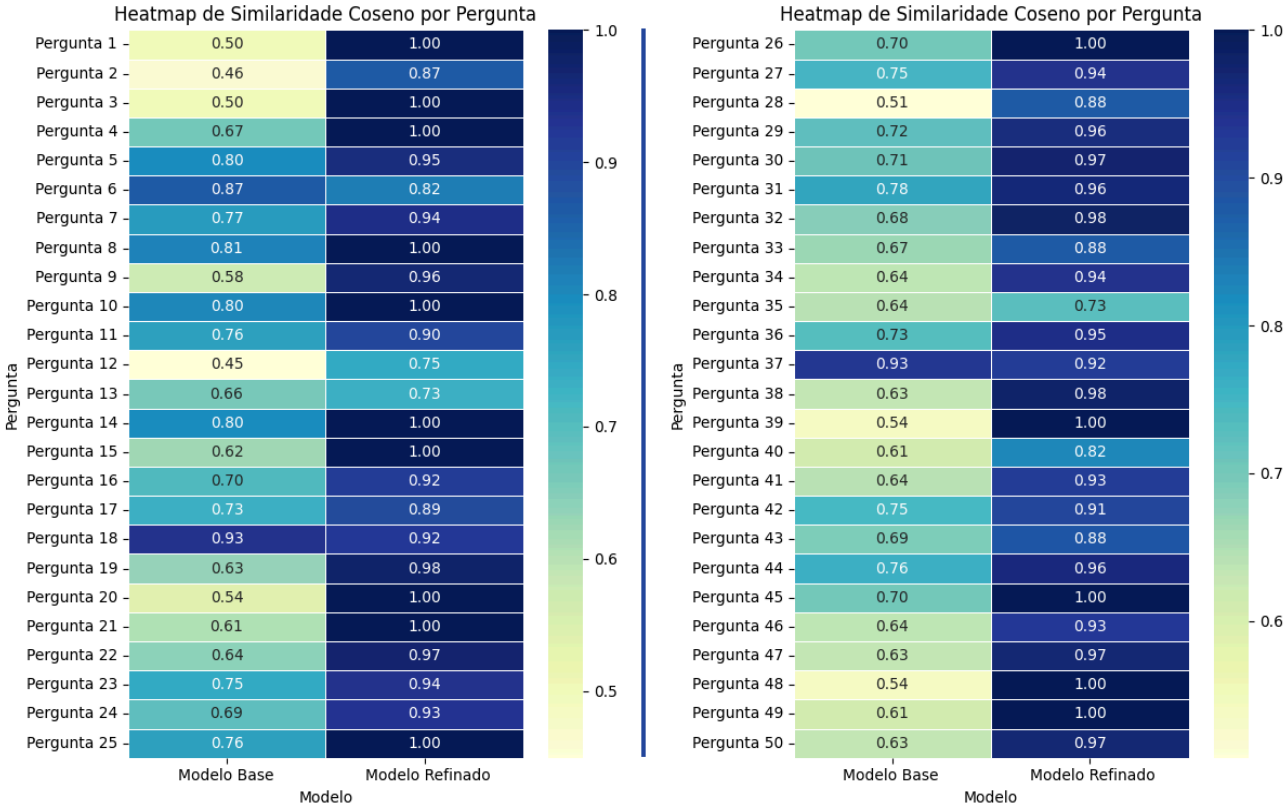
Fonte: Elaborada pelos autores.

4.3. AVALIAÇÃO DO APRIMORAMENTO

Por fim, foi realizada a avaliação da melhoria encontrada ao comparar as respostas resultantes coletadas dos modelos GPT-3.5 base e refinado. Para esta parametrização, foram geradas respostas para as mesmas perguntas utilizadas na primeira avaliação, realizando o envio das mesmas ao modelo base e coletando os resultados para comparação com as respostas ideais respectivas.

Visualizando os resultados obtidos, representados no gráfico exibido na Figura 10, é possível comparar diferenças de proximidade apresentadas entre as respostas do modelo base comparadas às respostas do modelo após o *fine-tuning*.

Figura 11 – Gráfico mapa de calor de comparação da similaridade considerando a base de respostas ideais e as respostas geradas pelo modelo *fine-tuning*, onde cada uma das partes apresenta metade do total de perguntas e seus respectivos valores na escala.



Fonte: Elaborada pelos autores.

A média resultante dos valores obtidos, também calculados através da utilização da métrica *Cosine Similarity*, indicou aproximadamente 0,678 de similaridade entre as respostas geradas pelo modelo base e as respostas ideais. Ao realizar a comparação deste resultado com o valor médio obtido ao mensurar utilizando as respostas geradas pelo modelo pós-treinamento, é observada uma melhoria de 27,8% em comparação com o resultado do modelo base.

Na Tabela 3, são exibidas amostras das respostas geradas pelo modelo base e refinado, comparadas com as respostas consideradas ideais incluídas na base. Nos exemplos da tabela, é possível perceber que as respostas da versão refinada se assemelham às respostas consideradas ideias.

Tabela 2 - Demonstração das respostas obtidas do modelo base e refinado.

Pergunta	Resposta ideal	Resposta do modelo GPT-3.5 refinado	Resposta do modelo GPT-3.5 base
Preciso de ajuda	Olá, tudo bem? \n Sou o assistente especializado em questões sobre cartões. Como posso te ajudar? \n - Cancelamento do cartão \n - Adicional de cartão \n - Limite do cartão	Olá, tudo bem? \n Sou o assistente especializado em questões sobre cartões. Como posso te ajudar? \n - Cancelamento do cartão \n - Adicional de cartão \n - Limite do cartão	Claro! Me diz com o que você precisa de ajuda — pode ser sobre código, design, texto, trabalho, estudo, vida pessoal… estou por aqui.
Quais são os benefícios de ser um sócio torcedor?	Como sócio torcedor, você tem acesso prioritário à compra de ingressos, descontos em produtos oficiais e participação em eventos exclusivos.	Como sócio torcedor, você garante acesso prioritário na compra de ingressos, aproveita descontos em produtos oficiais e ainda pode participar de eventos exclusivos preparados especialmente para você.	Os benefícios de ser um sócio torcedor geralmente incluem desconto em ingressos, acesso exclusivo a pré-venda, participação em promoções, brindes oficiais, e experiências exclusivas com o clube.
Como posso alterar a forma de pagamento do meu plano de sócio torcedor?	A alteração da forma de pagamento pode ser realizada acessando sua conta no site oficial, na área de sócios, ou entrando em contato com o setor financeiro do clube.	A alteração da forma de pagamento pode ser feita ao acessar sua conta na área de sócios do site oficial ou, se preferir, entrar em contato com o setor financeiro do clube.	A forma de pagamento pode ser modificada acessando sua conta no portal do sócio ou entrando em contato com o atendimento.
Existe um prazo para atualizar meus dados cadastrais no plano de sócio torcedor?	Não há um prazo específico para atualizar seus dados cadastrais, mas é importante manter suas informações sempre atualizadas para evitar problemas na comunicação e no uso dos benefícios do plano de sócio torcedor.	Não há um prazo específico para atualizar seus dados cadastrais, mas é sempre muito importante manter suas informações em dia para evitar problemas de comunicação e garantir o uso tranquilo de todos os benefícios do plano de sócio torcedor.	Não há um prazo fixo, mas é recomendado manter os dados atualizados para não perder benefícios.
Qual é o processo para se tornar sócio torcedor?	Para se tornar sócio torcedor, basta acessar o site oficial do clube, escolher o plano desejado, preencher o cadastro e realizar o pagamento online.	Para se tornar sócio torcedor, basta acessar o site oficial do clube, escolher o plano que mais combina com você, preencher o cadastro de forma simples e, por fim, realizar o pagamento online com segurança.	Basta se cadastrar no site oficial do programa, escolher um plano e realizar o pagamento.

4.4. ANÁLISE DE MÉTRICAS

Foi observado, considerando os resultados obtidos após a mensuração com base nas métricas utilizadas, que os dados retornados são consistentes e estão alinhados com os dados inseridos na base de treinamento, exibindo boa capacidade de

condução do modelo para com o cliente e da geração de respostas adequadas e satisfatórias.

Também foi observado que o custo é relativamente baixo para os modelos GPT da OpenAI, quando comparado a outros modelos de geração textual existentes no mercado, como o Watson, modelo da IBM. O Watson, além de cobrar o valor aproximado de US$0,0006 por 1000 *tokens*, exige uma assinatura mensal de US$1,050 para o plano Standard, transformando seu uso consideravelmente mais caro que os modelos da OpenAI (IBM, 2024a; IBM, 2024b).

A criação e implementação do modelo foram bem intuitivas, com documentação clara e explicativa sobre o passo a passo requerido em todas as etapas. Os resultados obtidos foram satisfatórios, as respostas da solução são rápidas para o propósito dela, e o custo é baixo para utilização. Apesar de desafios como a necessidade de reprocessar o treinamento em cada ajuste e a complexidade na integração com o sistema síncrono para finalização de interações com o cliente, o projeto contribui para o avanço de soluções de inteligência artificial aplicadas no atendimento, atendendo assim a demanda prevista.

5. CONCLUSÃO

Este projeto teve como objetivo apresentar a criação de um *chatbot* voltado à automatização e aprimoramento do processo de atendimento ao cliente, utilizando como ferramenta principal o ChatGPT junto a técnicas de *fine-tuning* aplicadas a seus modelos já existentes.

Este mecanismo permite o direcionamento do *bot* a temas específicos determinados de acordo com a necessidade da empresa utilizadora do recurso, determinando também o escopo da base de perguntas e respostas a ser utilizada, as quais foram extraídas de interações reais e representam as necessidades verdadeiras dos clientes.

Ao realizar a curadoria da base de perguntas e respostas a ser utilizada e efetuar o refinamento do modelo GPT-3.5, foram elaborados e executados testes envolvendo o tempo de resposta do modelo, a similaridade entre as respostas ideais e as geradas pelo modelo refinado e, por fim, o enquadramento das respostas no escopo considerando as respostas geradas pelo modelo com *fine-tuning* e as respostas geradas pelo modelo base, sendo as duas últimas situações mensuradas utilizando o cálculo de similaridade por cosseno.

Para o teste de tempo de resposta, o tempo médio do total, considerando a média calculada sobre todos os testes realizados, foi de 1,04 segundos, estando dentro do tempo ideal definido pela empresa solicitante.

Em relação à avaliação de similaridade das respostas, o modelo apresentou uma pontuação média de 0,939, conforme a métrica adotada. Além disso, ao analisar o enquadramento no escopo, comparando as respostas do modelo base com as do modelo refinado, observou-se uma melhoria de 27,8%. O modelo original havia obtido pontuação média de 0,678, enquanto o modelo treinado apresentou desempenho significativamente superior.

Esses resultados demonstram que o refinamento contribuiu de forma consistente para o alinhamento das respostas ao escopo definido, atendendo satisfatoriamente às necessidades estabelecidas no início do projeto.

5.1. LIMITAÇÕES

As limitações percebidas envolvem tanto o processo de treinamento quanto a etapa de integração. Uma das principais dificuldades está na impossibilidade de ajustar apenas partes específicas da base quando são identificados problemas de escopo ou respostas incoerentes.

Sempre que a base de perguntas e respostas precisa ser modificada — seja para corrigir inconsistências, ampliar o escopo ou substituir exemplos inadequados — é necessário realizar um novo processo completo de fine-tuning, utilizando novamente toda a base revisada.

Esse procedimento torna a correção e o aprimoramento do modelo mais custosos e demorados, já que demanda tempo computacional adicional e aumenta o esforço sempre que a base de dados sofre alterações, especialmente em bases de maior tamanho.

Foi constatado também que a preparação da plataforma apresentou desafios adicionais, uma vez que a troca síncrona de mensagens ocorre entre o usuário final (cliente da empresa parceira que utiliza o chat) e o CRM. A adaptação desse fluxo para comportar o funcionamento da API assíncrona da OpenAI responsável pela comunicação com o modelo treinado mostrou-se mais complexa do que o previsto.

Por este motivo, sendo necessária a implementação de um mecanismo que realiza a troca para a interação humana ou finalização da interação no sistema, foram adicionadas à base frases e perguntas-chave, cujas respostas retornavam palavras únicas que serviram como *tags* de identificação de gatilhos para interromper a conversa.

Este processo se mostrou trabalhoso, visto que por vezes as respostas retornadas não se tratavam unicamente das palavras registradas, incluindo frases ou até mesmo caracteres diferentes dos desejados. Em outros mecanismos de treinamento, o processo de integração é mais otimizado, pois permite a criação e retorno de variáveis voltadas à comunicação entre os sistemas.

5.2. TRABALHOS FUTUROS

Após a implementação com o modelo de *fine-tuning* do ChatGPT, surgiram novas oportunidades para aprimorar o sistema e buscar outros modelos de IA, como o Watson, da empresa IBM, e de plataformas que usam os próprios modelos GPT, como a ZAIA. Atualmente, já estão em andamento implementações e os resultados aparentam ser igualmente atrativos.

É importante destacar que, nas etapas iniciais do projeto, a utilização de um modelo de menor escala ou de alternativas open source poderia ter sido mais apropriada, considerando o tempo reduzido disponível para o desenvolvimento, a exigência por baixo custo e a necessidade de maior agilidade na implementação. Essa escolha teria tornado o processo mais acessível e diminuído o esforço de preparação da base de dados utilizada para o treinamento.

A IA generativa, apesar de poderosa, é mais indicada quando se tem uma base de conhecimento de proporções consideravelmente maiores e necessidade de comportamento altamente específico. Por fim, a construção deste trabalho resultou em uma integração com o sistema eficiente, mas com algumas ressalvas para melhorar futuramente. Além disso, representou uma experiência enriquecedora, proporcionando o aprofundamento em metodologias e tecnologias amplamente utilizadas no mercado de trabalho.

6. REFERENCIAS

GORDON, I. *Marketing de relacionamento: estratégias, técnicas e tecnologia para conquistar clientes e mantê-los para sempre*. São Paulo: Futura: 1999. Acesso em: 01 dez. 2024.

FRIEDMAN, Thomas L. *O mundo é plaPricing no: uma breve história do século XXI*. São Paulo: Objetiva, 2005. Acesso em: 01 dez. 2024.

TORRES, Claudio. *A bíblia do marketing digital: tudo o que você queria saber sobre marketing e publicidade na internet e não tinha a quem perguntar.* São Paulo: Novatec Editora, 2009. Acesso em: 13 jan. 2025.

WHATSAPP no Brasil: conheça a trajetória do app. Blip Blog, 23 nov. 2022. Disponível em: https://www.blip.ai/blog/whatsapp/whatsapp-no-brasil/. Acesso em: 11 fev. 2025.

MATTAR, Auana. IA no atendimento ao cliente: como potencializar o uso da tecnologia para entregar experiências mais fluidas. *MIT Technology Review Brasil*. Disponível em: https://mittechreview.com.br/ia-atendimento-cliente/. Acesso em: 26 nov. 2024.

ADAMOPOULOU, Eleni; MOUSSIADES, Lefteris. Chatbots: History, technology, and applications. Machine Learning with applications, 2020. Disponível em: https://doi.org/10.1016/j.mlwa.2020.100006

JUNIPER RESEARCH. Retail Spend Over Chatbots to Reach $12bn Globally by 2023. Disponível em: https://www.juniperresearch.com/press/retail-spend-over-chatbots-to-reach-12bn-globally/. Acesso em: 26 nov. 2024.

DEMANDSAGE. *ChatGPT Statistics (NOV. 2024) – 200 Million Active Users.* Disponível em: https://www.demandsage.com/chatgpt-statistics/. Acesso em: 26 nov. 2024.

QIAN, Lihong; WANG, I. Kim. Competition and innovation: The tango of the market and technology in the competitive landscape. Managerial and Decision Economics, 2017. Disponível em: https://doi.org/10.1002/mde.2861. Acesso em: 12 ago. 2025.

PATIL, Dimple; DIGITAL Hurix; INDIA, Andheri. Artificial intelligence-driven customer service: Enhancing personalization, loyalty, and customer satisfaction. Loyalty, And Customer Satisfaction, 2025. Disponível em: https://dx.doi.org/10.2139/ssrn.5057432. Acesso em: 12 ago. 2025.

NICOLESCU, Luminita; TUDORACHE, Monica. Human-Computer Interaction in Customer Service: The Experience with AI Chatbots — A Systematic Literature Review. Electronics, 15

mai. 2022. Disponível em: https://doi.org/10.3390/electronics11101579. Acesso em: 12 ago. 2025.

BOCCHI, Rodrigo. Prefácio. In: MUNIZ, A.; SOUZA, A. G. de; GRELLT, B. M.; LABRIOLA, L. E. *Jornada da experiência do cliente: unindo práticas e metodologias da cultura customer centric para alcançar crescimento e geração de resultados com foco no cliente!*. Rio de Janeiro: Brasport, 2022. p. 5.

JENA, Suvendu; YANG, Jilei; TAN, Fangfang. Unlocking Sales Growth: Account Prioritization Engine with Explainable AI. arXiv e-prints, 12 jun. 2023. Disponível em: arXiv:2306.07464. Acesso em: 22 jul. 2025.

MCKINSEY & COMPANY. *The state of AI: The state of AI in 2021*. 2021. Disponível em: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai. Acesso em: 20 fev. 2025.

SHIHAB, Khalil. A Backpropagation Neural Network for Computer Network Security. Journal of Computer Science 2.9, 2006. Disponível em: https://doi.org/10.3844/jcssp.2006.710.715. Acesso em 6 set. 2025.

ZHAO, Wayne Xin; ZHOU, Kun; LI, Junyi; et al. A survey of large language models. arXiv preprint arXiv:2303.18223, v.16, 2023. Disponível em: https://doi.org/10.48550/arXiv.2303.18223. Acesso em: 6 set. 2025.

NAVEED, H.; KHAN, A. U.; QIU, S.; SAQIB, M.; ANWAR, S.; USMAN, M.; AKHTAR, N.; BARNES, N.; MIAN, A. A Comprehensive Overview of Large Language Models. Disponível em: arXiv. Preprint, Ceará, 12 jul. 2023. Disponível em: https://doi.org/10.48550/arXiv.2307.06435. Acesso em: 6 set. 2025.

WEI, Jason; TAY Yi; BOMMASANI, Rishi; RAFFEL Colin; ZOPH, Barret; BORGEAUD, Sebastian; YOGATAMA, Dani; BOSMA, Maarten; ZHOU, Denny; METZLER, Donald; CHI, Ed H; HASHIMOTO, Tatsumori; VINYALS, Oriol; LIANG, Percy; DEAN, Jeff; FEDUS, William. Emergent Abilities of Large Language Models. arXiv, 2022. Disponível em: https://doi.org/10.48550/arXiv.2206.07682. Acesso em: 6 set. 2025.

SUSNJAK, Teo. ChatGPT: The End of Online Exam Integrity? arXiv, 2022. Disponível em: https://doi.org/10.48550/arXiv.2212.09292. Acesso em: 18 ago. 2025.

ILSE, Benjamin; BLACKWOOD, Frederick. Comparative Analysis of Finetuning Strategies and Automated Evaluation Metrics for Large Language Models in Customer Service Chatbots. 2024. Disponível em: https://doi.org/10.21203/rs.3.rs-4895456/v1. Acesso em: 18 nov. 2025.

YUHAO, Dan et al. EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. 2023. Disponível em: https://doi.org/10.48550/arXiv.2308.02773. Acesso em: 17 nov. 2025.

LI, Yiming et al. VaxBot-HPV: a GPT-based chatbot for answering HPV vaccine-related questions. JAMIA open, 2024. Disponível em: https://doi.org/10.1093/jamiaopen/ooaf005. Acesso em 17 nov. 2025.

TZANIS, Nikolaos. Creation of a chatbot using language models and deep learning for customer question answering. Dissertação de Mestrado. UNIVERSITY OF PIRAEUS, 2025. Disponível em: https://dione.lib.unipi.gr/xmlui/handle/unipi/17742. Acesso em: 19 nov. 2025.

OPENAI**.** *Pricing*. Disponível em: https://openai.com/api/pricing/. Acesso em: 05 jan. 2025.

IBM. *watsonx.ai – Pricing*. Disponível em: https://www.ibm.com/products/watsonx-ai/pricing. Acesso em: 06 jun. 2025.

IBM. *Billing details for generative AI assets*. Disponível em: https://www.ibm.com/docs/en/watsonx/saas?topic=plans-billing-details-generative-ai-assets. Acesso em: 06 jun. 2025.