FUNDAÇÃO EDUCACIONAL DE FERNANDÓPOLIS
BACHARELADO EM SISTEMAS DE INFORMAÇÃO

CHATBOT PARA ATENDIMENTO DE CHAMADO INICIAL DE T.I
CHATBOT FOR FIRST-LEVEL IT SUPPORT TICKET MANAGEMENT

Lucas dos Santos Uvera[1], Guilherme Vieira de Camargo[2], Jefferson Antonio Ribeiro Passerini[3]

[1]Fundação Educacional de Fernandópolis, lucasuvera@gmail.com
[2] Fundação Educacional de Fernandópolis, camargoguilhermed99@gmail.com
[3] Fundação Educacional de Fernandópolis, jefferson.passerini@fef.edu.br

Área: Ciência de dados e Inteligência artificial
Subárea: Modelagem e controle de processos

RESUMO

O avanço da tecnologia tem impulsionado a adoção de soluções automatizadas nos mais diversos setores, incluindo o suporte técnico em Tecnologia da Informação (TI). Entre essas soluções, os *chatbots* têm ganhado destaque por sua capacidade de realizar atendimentos iniciais de forma rápida, padronizada e eficiente. Este trabalho apresenta o desenvolvimento de um *chatbot* voltado especificamente para o atendimento inicial de chamados de TI, com foco na triagem de solicitações, resposta a dúvidas frequentes e encaminhamento adequado para os setores responsáveis. Diferentemente dos sistemas tradicionais baseados em fluxos rígidos de decisão, o projeto utiliza modelos de linguagem de larga escala (*LLMs*), como o *GPT-3*, que possibilitam maior flexibilidade na compreensão da linguagem natural dos usuários. A proposta visa integrar inteligência artificial ao ambiente corporativo, tornando o processo de atendimento mais dinâmico e reduzindo a sobrecarga sobre as equipes humanas de suporte. A pesquisa busca demonstrar a viabilidade técnica e prática da utilização de *LLMs* nesse contexto, destacando benefícios como a redução do tempo médio de resposta, aumento da satisfação dos usuários e otimização dos custos operacionais. Os resultados esperados incluem a melhoria da eficiência no atendimento de TI e o avanço na aplicação de tecnologias baseadas em IA no setor corporativo.

Palavras chaves: Chatbot; Inteligência Artificial; Automação; LLM; Llama 3.

ABSTRACT

The advancement of technology has driven the adoption of automated solutions across various sectors, including technical support in Information Technology (IT). Among these solutions, chatbots have stood out for their ability to perform initial customer service quickly, consistently, and efficiently. This work presents the development of a chatbot specifically designed for the initial handling of IT support tickets, focusing on issue triage, frequently asked questions, and proper routing to the responsible departments. Unlike traditional systems based on rigid decision flows, the project uses Large Language Models (LLMs), such as GPT-4, which allow greater flexibility in understanding users' natural language. The proposal aims to integrate artificial intelligence into corporate environments, making the support process more dynamic and reducing the burden on human support teams. This research seeks to demonstrate the technical and practical feasibility of using LLMs in this context, highlighting benefits such as

reduced average response time, increased user satisfaction, and operational cost optimization. The expected outcomes include improved efficiency in IT support services and progress in applying AI-based technologies in the corporate sector.

Keywords: *Chatbot; IT Support; Artificial Intelligence; Language Models; Automation; Ollama*

1 INTRODUÇÃO

A implementação de chatbots no atendimento de TI tem se mostrado uma estratégia eficaz para aumentar a eficiência operacional e diminuir o tempo de resposta aos usuários.

No entanto, muitos desses sistemas ainda funcionam com fluxos de arvores de decisão, o que limita sua capacidade de interpretar a linguagem natural e de se adaptar a diferentes formas de comunicação utilizadas no suporte técnico.

Com o surgimento de modelos de linguagem de larga escala, os chamados *Large Language Models (LLMs),* como o *GPT-3*, tornou-se possível construir assistentes conversacionais capazes de interpretar comandos com maior flexibilidade e responder de forma contextualizada e relevante (BROWN et al., 2020).

Nesse cenário, o presente trabalho propôs o desenvolvimento de um chatbot para atendimento inicial de chamados de TI utilizando um LLM, com foco na triagem de problemas, resposta a dúvidas frequentes e encaminhamento adequado aos setores responsáveis.

O projeto busca comprovar a viabilidade técnica e prática da aplicação de LLMs em ambientes corporativos, visando automatizar tarefas repetitivas, reduzir custos operacionais e melhorar a experiência do usuário no atendimento técnico.

Como resultado central deste projeto, a implementação piloto demonstrou a viabilidade técnica e prática da utilização de *LLMs open-source* em infraestrutura local *(on-premise)*. Os dados coletados evidenciam que, ao eliminar a dependência de *APIs* externas, a solução não apenas garantiu a privacidade absoluta dos dados corporativos, mas também obteve altos índices de aceitação por parte dos usuários devido à agilidade das respostas. A validação prática confirmou que o uso de modelos como o *Llama* 3, integrados a uma base de conhecimento contextual, supera as limitações dos fluxos rígidos tradicionais, oferecendo uma experiência de autoatendimento mais fluida e eficiente para o suporte de TI.

2 REFERENCIAL TEÓRICO

Essa seção tem como objeto esclarecer e mostrar os conceitos e metodologia proposta, abrangendo as informações referente a inteligencia artificial e *LLM,* bem como modelos e algoritmos apresentados neste trabalho.

2.1 A Evolução do Suporte de TI e a Ascensão dos Chatbots

A criação dos *chatbots* pode estar ligada na visão de Alan Turing, mais especificamente do “Teste de Turing”, também chamado de “Jogo da Imitação”. Criado em 1949, é um teste da capacidade da máquina de exibir comportamento inteligente equivalente

ao de um ser humano. No teste, o avaliador julga uma transcrição de texto de uma conversa em linguagem natural entre um ser humano e uma máquina, e caso o avaliador não identificar a máquina a mesma passa. O teste não tinha como objetivo identificar se a máquina conseguia responder as perguntas corretamente, e sim se a suas respostas estavam semelhantes à de um humano.

Seu teste foi introduzido em 1950 no artigo "*Computing Machinery and Intelligence*". No entanto, para compreender a profundidade de sua proposta, é crucial entender a questão que ele procurava contornar. Turing considerava a pergunta "Podem as máquinas pensar?" como "demasiado sem sentido para merecer discussão" (Turing, 1950, p. 442). O problema reside na ambiguidade das palavras "máquina" e "pensar". Definir esses termos de forma satisfatória levaria a um debate mais filosófico do que científico.

O trabalho de Alan Turing estabeleceu a base teórica para o campo da Inteligência Artificial. A ideia de julgar a inteligência de uma máquina com base em sua performance conversacional é o princípio fundamental por trás de todos os chatbots modernos.

Tendo sua criação em 1962, dezenove anos depois do início dos estudos sobre Inteligência Artificial. O primeiro programa de grande repercussão foi o ELIZA, criado por Joseph Weizenbaum no MIT entre 1964 e 1966. ELIZA simulava uma conversa com uma psicoterapeuta rogeriana. Seu método era engenhosamente simples: ela não entendia a conversa, mas utilizava um sistema de reconhecimento de padrões e substituição de palavras-chave. Por exemplo, se um usuário digitasse "Eu estou triste com meu pai", ELIZA poderia identificar o padrão "meu X" e responder com uma frase pré-programada como "Fale-me mais sobre seu X", resultando em "Fale-me mais sobre seu pai".

Tendo em vista este exemplo, podemos constatar que, *chatbots* podem ser aplicados em inúmeras áreas da sociedade, trazendo benefícios em todos os segmentos em que se é aplicado, como maior eficiência de atendimento, melhor distribuição de tempo para atendentes reais e maior lucro para empresas que a utilizam.

De acordo com Suhaili, Sinarwati Mohamad, Naomie Salim, e Mohamad Nazim Jambli (2021), *chatbots* representam o próximo salto tecnológico significativo no campo dos serviços conversacionais, ou seja, permitir que um dispositivo se comunique com um usuário ao receber solicitações em linguagem natural. O dispositivo utiliza inteligência artificial e aprendizado de máquina para responder ao usuário com respostas automatizadas. Embora esta seja uma área de estudo relativamente nova, a aplicação deste conceito aumentou substancialmente nos últimos anos. A tecnologia não se limita mais a emular conversas humanas, mas também está sendo cada vez mais utilizada para responder a perguntas, seja em ambientes acadêmicos ou em usos comerciais, como situações que exigem que os assistentes busquem razões para a insatisfação do cliente ou recomendem produtos e serviços.

A área de Tecnologia da Informação (TI) é o sistema nervoso de qualquer organização no cenário atual. Sua eficiência impacta diretamente a produtividade de todos os outros setores e o funcionamento da própria empresa. No entanto o atendimento de chamados de TI, com a utilização de ferramentas também conhecidos como *Help Desks* ou *Service Desks* (Suporte Técnico), enfrentam um desafio constante: um alto volume de solicitações, muitas das quais são repetitivas e de baixa complexidade, como redefinições de senha, problemas de conexão, instalação de softwares padronizados e dúvidas sobre o uso de sistemas. Este

cenário cria um ambiente ideal para a aplicação de automação inteligente, onde os *chatbots* evoluíram de meras ferramentas de conveniência para componentes estratégicos da operação.

A implementação de *chatbot* inteligente no suporte de TI não é apenas uma modernização tecnológica. É uma decisão estratégica com impactos profundos de eficiência, custos e satisfação do usuário.

2.2 Utilização de LLMs no ambiente Corporativo

A adoção de *LLMs* no ambiente corporativo se divide em duas grandes abordagens. De um lado, modelos proprietários, como o *GPT-4 (OpenAI)* e o *Claude 3 (Anthropic),* oferecem desempenho de ponta através de *APIs* pagas. Esta abordagem, embora de fácil implementação, transfere o processamento dos dados para servidores externos, criando preocupações legítimas sobre a confidencialidade de informações estratégicas ou dados de clientes.

Por outro lado, o movimento de código aberto *(open-source)* tem ganhado tração, impulsionado por modelos que rivalizam em capacidade com suas contrapartes fechadas. A disponibilização de *LLMs* de código aberto permite que as organizações os modifiquem, ajustem (façam *fine-tuning*) e, o mais importante, os executem em sua própria infraestrutura.

2.3 Llama 3

Lançado pela Meta, o Llama 3 é a mais recente geração da família de modelos de linguagem de código aberto. Disponível em diferentes tamanhos (como 8B e 70B parâmetros), o Llama 3 foi treinado em um conjunto massivo de mais de 15 trilhões de tokens. Isso resultou em capacidades significativamente melhoradas de raciocínio, geração de código e, crucialmente para este projeto, compreensão e geração de diálogo em múltiplos idiomas, incluindo o português. Sua arquitetura aberta o torna ideal para customização e implementação em casos de uso empresariais.

2.4 Configuração para execução local do llama 3

Para que um modelo como o Llama 3 seja funcional em um ambiente corporativo, é necessária uma plataforma para gerenciá-lo. O Ollama é uma ferramenta de código aberto projetada para simplificar o download, a configuração e a execução de LLMs em hardware local, seja em um servidor ou em uma estação de trabalho.

A principal vantagem estratégica do Ollama é a privacidade. Ao utilizar o Ollama, o chatbot processa a consulta do usuário (ex: "Estou com problema de acesso ao servidor X") inteiramente dentro da rede local da empresa. Nenhuma informação sensível é enviada pela internet, mitigando riscos de vazamento de dados e garantindo conformidade com regulações de privacidade. Além disso, a execução local elimina custos variáveis de API e reduz a latência, resultando em respostas quase instantâneas.

2.5 Engenharia de *Prompt* e *In-Context Learning*

O modelo Llama 3, apesar de ser um Modelo de Linguagem de Larga Escala (LLM) de alta capacidade, não possui conhecimento prévio sobre os procedimentos, sistemas ou políticas internas de uma organização específica. Para que o chatbot de TI seja eficaz, deve-se operar com base em um conjunto de informações precisas e controladas, evitando a geração de respostas genéricas ou incorretas (fenômeno também conhecido como "alucinação").

Diferentemente de abordagens que exigem o complexo e custoso processo de fine-tuning (re-treinamento do modelo), este projeto emprega a técnica de Engenharia de Prompt combinada com o Aprendizado por Contexto (*In-Context Learning* - ICL).

A Engenharia de Prompt é o processo de desenhar e otimizar as instruções (*prompts*) fornecidas ao LLM para direcionar seu comportamento e suas respostas. Em vez de permitir uma conversação aberta, o *prompt* é estruturado para restringir e informar o modelo.

Conforme implementado no projeto, a estruturação do prompt enviado à API do Ollama baseia-se em dois componentes fundamentais. O primeiro é o Prompt de Sistema (*System Prompt*), responsável por definir a "persona" e as regras operacionais, estabelecendo o tom e o domínio de atuação do modelo através de instruções direcionadas, como "Você é um assistente de T.I. prestativo". O segundo componente é a Base de Conhecimento Contextual, que constitui o núcleo do Aprendizado por Contexto (*ICL*). Nesta etapa, a base de dados contendo problemas e soluções de TI é injetada dinamicamente no *prompt* junto à pergunta do usuário, acompanhada de instruções explícitas para que o modelo utilize exclusivamente as informações fornecidas no contexto para formular sua resposta.

A adoção dessa abordagem proporciona vantagens significativas para o ambiente corporativo. Em termos de agilidade, permite que a base de conhecimento seja atualizada em tempo real — como na adição de novos procedimentos — sem a necessidade de re-treinar ou reiniciar o modelo. Além disso, assegura maior controle e precisão, garantindo que o chatbot responda com base em informações verificadas e aprovadas pela empresa, o que eleva a confiabilidade das interações. Por fim, destaca-se a eficiência de recursos, uma vez que o método evita os elevados custos computacionais associados ao processo de fine-tuning de LLMs de código aberto.

2.6 Métricas de Avaliação para Chatbots de Suporte

A avaliação da eficácia do sistema baseou-se, primordialmente, na Taxa de Resolução no Primeiro Contato (First Contact Resolution Rate). Esta é considerada a métrica de eficácia mais crucial , pois mede a porcentagem de interações em que o chatbot conseguiu resolver o problema ou responder à dúvida do usuário de forma completa, eliminando a necessidade de escalonar o chamado para um analista humano. Paralelamente, foram analisadas as métricas de percepção e usabilidade, com foco na Facilidade de Uso (Ease of Use), que avalia a experiência do usuário (UX) e a simplicidade da interação , partindo da premissa de que um sistema tecnicamente perfeito terá baixa adoção se for difícil de operar.

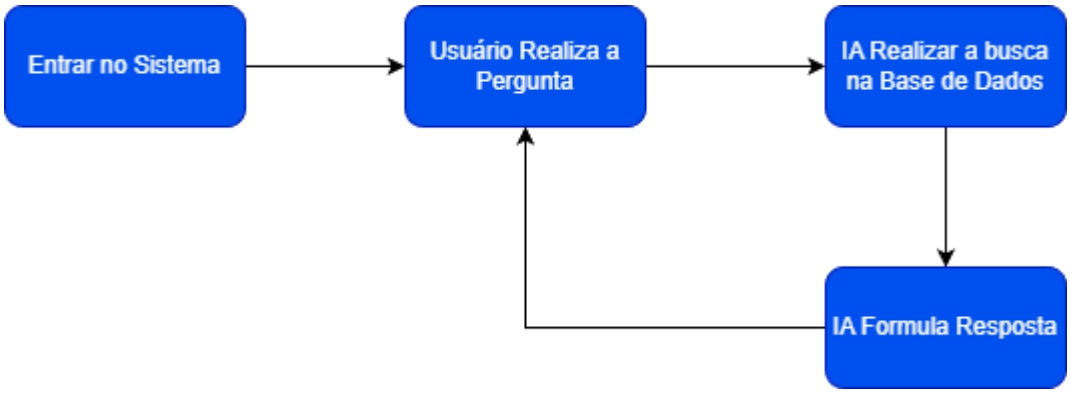
Para a mensuração desses aspectos subjetivos, utilizou-se um formulário de coleta que aplicou o Net Promoter Score (NPS) para quantificar o quanto o usuário gostou do chatbot , em conjunto com o Customer Satisfaction Score (CSAT), que avalia a satisfação geral com a qualidade da resposta e do atendimento. Adicionalmente, aferiu-se a Intenção de Adoção (Adoption Intent), métrica que define a probabilidade de o usuário voltar a utilizar a ferramenta, servindo como um forte indicador de confiança e valor percebido. Por fim, a análise foi complementada pelo Feedback Qualitativo, obtido através de comentários abertos para capturar nuances que escapam aos dados quantitativos, sendo essencial para identificar

falhas específicas e sugestões de melhoria. A combinação destas métricas permitiu uma análise holística, avaliando não apenas a funcionalidade técnica, mas também a eficiência e a agradabilidade da experiência para os usuários finais.

3 METODOLOGIA

Este projeto configura-se como uma pesquisa aplicada de natureza mista (quantitativa e qualitativa). O objetivo prático é o desenvolvimento de um protótipo funcional, o *chatbot*, para resolver o problema real da triagem de chamados de TI, sendo a abordagem quantitativa utilizada na análise dos dados do formulário de satisfação (taxa de resolução, facilidade de uso) e a qualitativa na análise dos comentários abertos dos usuários. A solução foi projetada sob o paradigma *on-premise* (localmente), garantindo que nenhuma informação saia da rede local. A arquitetura da solução, conforme exemplificado na Figura 1, é composta por uma Interface de Usuário (*Frontend*) desenvolvida em *Python* com a biblioteca *Streamlit*, uma Aplicação (*Backend*) no próprio código em *Python* que gerencia a conversa e a engenharia de prompt, e um Servidor do Modelo (*LLMaaS* Local), o *Ollama*, servindo o modelo *Llama* 3. O fluxo de interação se inicia na interface *Streamlit*, passa pelo *backend Python* que formata o *prompt* e chama a API local do *Ollama,* que por sua vez processa a requisição com o *Llama* 3 e retorna a resposta para ser exibida na interface.

Figura 1 – Arquitetura da chatbot *On-premise*.



Fonte: Elaborada pelos autores.

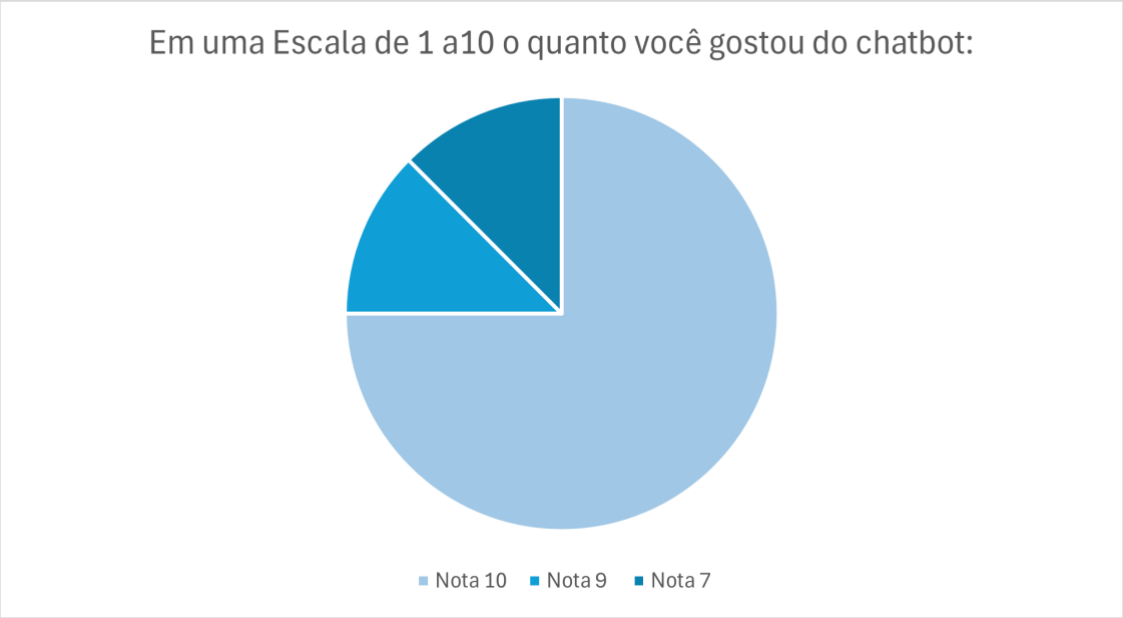
A escolha da linguagem de programação *Python* deve-se à sua vasta biblioteca de ferramentas para IA, enquanto o *Streamlit* foi selecionado pela sua capacidade de criar rapidamente interfaces *web* interativas. O uso do *Ollama* e do *Llama 3* (versão 8B) foi uma decisão estratégica para garantir a privacidade dos dados em execução via *localhost* e utilizar um modelo de código aberto de alta performance, comunicando-se via *API REST/HTTP*.

O desenvolvimento foi executado em três fases. Primeiramente, na Fase 1, realizou-se a configuração do ambiente local, com a instalação do *Ollama* e o *download* do modelo *Llama 3 (ollama pull llama3)*. Na Fase 2, foi feita a coleta e estruturação da Base de Conhecimento, colentando POPs (Procedimento Operacional Padrão, normativos e manuais intrenos, levantando os problemas e solicitações mais frequentes do setor de feitos pelo setor administrativo (alvo do piloto) e incorporando-os diretamente no projeto em *Python*, para instruir o modelo via Aprendizado por Contexto (Seção 2.5). Por fim, na Fase 3, desenvolveu-se a aplicação em *Streamlit*,responsável por inicializar a interface, gerenciar o histórico da

conversa, formatar o *prompt* final e exibir a resposta do *Llama* 3. Para a avaliação do protótipo, foi conduzido um projeto piloto com usuários do setor administrativo da minha empresa. O instrumento de coleta foi um formulário feito no *Google Forms*, aplicado após a interação com o chatbot, para aferir as métricas definidas na Seção 2.6: Pontos Positivos (campo aberto), Pontos Negativos (campo aberto), Satisfação/CSAT (NPS 1-10) e Feedback Qualitativo (campo aberto).

A análise dos dados, apresentada no Capítulo 4, seguirá uma abordagem quantitativa, calculando percentuais e médias das respostas fechadas, e uma abordagem qualitativa (Análise de Conteúdo) para agrupar e interpretar os comentários abertos. A Figura 2 exemplifica o tipo de análise quantitativa realizada sobre os dados coletados no piloto, demonstrando a Taxa de Resolução do chatbot.

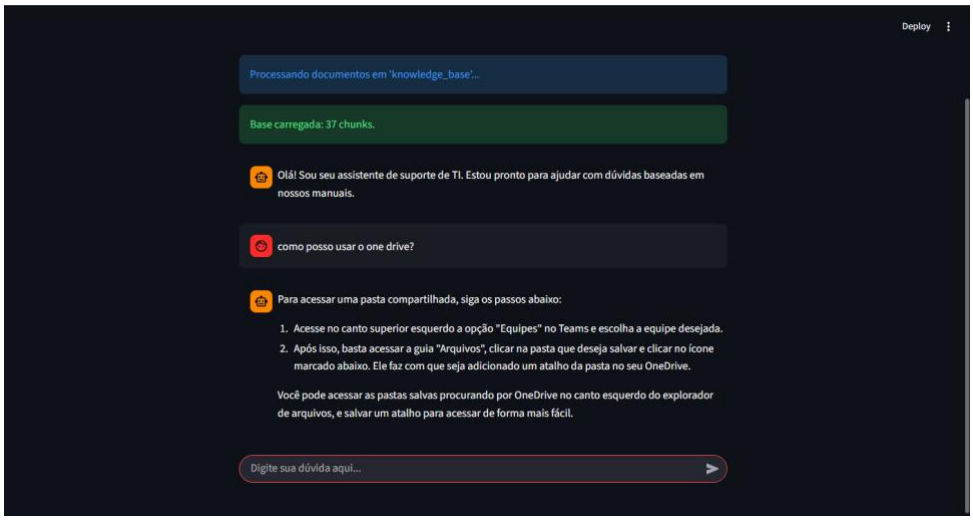
Figura 2 – Exemplo de análise da Taxa de Resolução (Resultados do Piloto)



Fonte: Elaborada pelos autores.

Demonstrando o funcionamento da solução proposta e ilustrar a experiência oferecida aos participantes do piloto, a figura a seguir apresenta um exemplo real de interação com o chatbot. A imagem demonstra a interface gráfica desenvolvida, evidenciando o ciclo completo de atendimento: a entrada da solicitação do usuário em linguagem natural e a resposta técnica estruturada gerada localmente pelo modelo, validando a integração entre os componentes da arquitetura descrita.

Figura 3 – Exemplo de interação do usuário com o *chatbot*



Fonte: Elaborada pelos autores.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Como resultado da pesquisa de *feedback* do *chatbot,* obteve-se uma amostra de *feedback* de 8 usuários. Os resultados foram obtidos através do formulário de avaliação, conforme detalhado na Metodologia, e são cruciais para validar a eficácia e a aceitação do piloto.

Os dados coletados do formulário são apresentados de forma consolidada nas tabelas e gráficos a seguir, refletindo as métricas centrais de desempenho e percepção do usuário.

A avaliação quantitativa da satisfação dos usuários é apresentada na Tabela 1. Os dados, mensurados através da métrica NPS (*Net Promoter Score*), revelam uma adesão massiva à solução proposta, com a predominância absoluta da pontuação máxima (10) e uma média geral elevada. Esses números validam a hipótese de que a interface simplificada e a proposta de autoatendimento foram bem recebidas pelo público-alvo do piloto.

Tabela 1 - Avaliação NPS do chatbot

Em uma escala de 1 a 10 o quanto você gostou do chatbot:
10
9
10
10
10
10
10
7

Fonte: Elaborada pelos autores.

No que tange aos aspectos qualitativos, a Tabela 2 compila os principais pontos de destaque percebidos pelos usuários. A análise das respostas abertas evidencia que a 'agilidade' e a 'facilidade de uso' foram os fatores mais elogiados. Isso corrobora a eficiência da arquitetura local (on-premise), que eliminou a latência de rede típica de serviços em nuvem, entregando respostas consideradas rápidas e objetivas pelos participantes.

Tabela 2 - Aspectos positivos percebidos

O que você mais gostou:
Respostas rápidas e precisas.
Respostas claras e objetivas.
Da facilidade ao acesso das respostas!
Fácil utilização
Velocidade de resposta
A rapidez da resposta
Mesmo sem a informação na base de dados, forneceu uma resposta lógica
A praticidade na procura de informações

Fonte: Elaborada pelos autores.

Por outro lado, as limitações da versão atual do protótipo estão expostas na Tabela 3. Embora alguns usuários não tenham identificado pontos desfavoráveis, as críticas existentes concentraram-se em dois eixos principais: o tempo de processamento da inferência em momentos específicos e, fundamentalmente, a restrição do escopo informacional. Esses feedbacks indicam que, apesar de funcional, o sistema necessita de uma base de conhecimento mais robusta para cobrir a variabilidade de dúvidas do setor.

Tabela 3 - Aspectos negativos identificados

O que você não gostou:
Não identifiquei nenhum ponto que seja desfavorável.
.
Do tempo de espera para o retorno da resposta.
Demorou um pouco para dar resposta
Sem resposta
Sem indicação
Sem observações
Precisar ter mas informações ampla para responder as perguntas

Fonte: Elaborada pelos autores.

Por fim, a Tabela 4 consolida as sugestões de melhoria oferecidas pelos participantes, apontando diretrizes claras para a evolução do projeto. As recomendações convergem para a necessidade de ampliar a base de dados e aumentar a capacidade generativa da Inteligência Artificial, sugerindo que os usuários desejam uma ferramenta ainda mais autônoma e abrangente para lidar com questões complexas que excedam o treinamento inicial.

Tabela 4 – Sugestões de melhorias do *chatbot*

O que poderia melhorar no chatbot?
Esta ótimo, funcional e ágil
Torne ele uma IA generativa.
Ampliação da base de dados
Sem indicação
Sem observações
Mais conhecido amplo

Fonte: Elaborada pelos autores.

Além dos dados numéricos, a análise qualitativa, obtida através do campo de comentários abertos do formulário, forneceu um contexto indispensável para compreender a percepção dos usuários. Os feedbacks foram agrupados em três temas centrais. Primeiramente, Pontos Fortes, onde foram recorrentes os elogios à rapidez, disponibilidade e facilidade de uso da ferramenta. Em segundo lugar, os Pontos de Melhoria (que justificam as falhas de resolução) centraram-se, de forma unânime, na limitação do escopo de conhecimento do chatbot .

A análise dos resultados, fundamentada na triangulação dos dados quantitativos e qualitativos obtidos por meio do formulário de satisfação dos usuários, revela uma avaliação amplamente positiva do chatbot desenvolvido. Os dados quantitativos mostram que a média geral de satisfação ficou próxima da pontuação máxima, com notas variando entre 7 e 10, evidenciando um alto nível de aceitação e aprovação por parte dos participantes (Tabela 1).

A análise qualitativa reforça essa percepção. Entre os principais aspectos positivos mencionados, destacam-se a rapidez e precisão das respostas, a facilidade de uso e a clareza das informações apresentadas. Esses elementos indicam que o *chatbot* atendeu de forma eficiente às necessidades básicas de atendimento, demonstrando bom desempenho técnico e usabilidade satisfatória. Por outro lado, as críticas identificadas foram pontuais, relacionadas principalmente ao tempo de resposta em algumas interações e à limitação da base de dados, o que aponta oportunidades de aprimoramento no escopo informacional do sistema. As sugestões de melhoria concentraram-se na ampliação da base de conhecimento e na evolução do chatbot para um modelo de inteligência artificial mais generativo e abrangente.

De modo geral, os resultados validam a proposta do projeto, indicando que o chatbot apresenta potencial real de aplicação prática e boa aceitação pelos usuários finais, com pontos de melhoria claros e de fácil implementação para versões futuras.

Esta aparente contradição – onde usuários cujo problema não foi resolvido ainda pretendem usar a ferramenta – encontra respaldo em estudos sobre o Modelo de Aceitação de Tecnologia (TAM). O TAM postula que a intenção de usar uma tecnologia é determinada por sua 'Facilidade de Uso Percebida' e 'Utilidade Percebida'. Os resultados confirmam o primeiro pilar. Para o segundo pilar, os dados qualitativos sugerem que os usuários encontraram 'Utilidade' não apenas na resolução, mas também na imediatez da resposta e na confiança gerada pela privacidade. A escolha metodológica da arquitetura *on-premise* (*Ollama* + *Llama 3*) é fundamental aqui; a execução local elimina a latência da rede e, como apontado por um usuário, garante a soberania dos dados. Os usuários demonstraram preferir a agilidade e a segurança de um autoatendimento instantâneo, mesmo que a primeira tentativa falhe, em vez de esperar em canais de atendimento humanos.

5 CONSIDERAÇÕES FINAIS

O objetivo geral deste projeto foi desenvolver e avaliar um protótipo de chatbot para o atendimento inicial de chamados de TI, com o diferencial estratégico de utilizar um Modelo de Linguagem de Larga Escala (LLM) de código aberto (*Llama* 3) e executa-lo localmente. A proposta buscou endereçar desafios críticos de implementações de IA no ambiente corporativo, notadamente a privacidade de dados, o controle de infraestrutura e os custos operacionais associados a APIs proprietárias.

A metodologia de pesquisa aplicada, que culminou no desenvolvimento de um protótipo funcional e em um projeto piloto com 8 usuários de um setor administrativo, permitiu validar a tese central do trabalho. Os resultados do piloto demonstraram a viabilidade técnica da solução on-premise. O chatbot alcançou uma taxa de resolução no primeiro contato positivo provando ser eficaz para a maioria das interações.

Mais significativamente, o estudo revelou uma altíssima aceitação por parte dos usuários finais. Mesmo com uma taxa de resolução imperfeita, o protótipo obteve uma nota média de 9.5 (de 10). A análise qualitativa dos feedbacks esclareceu essa aparente contradição: os usuários valorizaram imensamente a disponibilidade imediata, a velocidade de resposta (proporcionada pela execução local via *Ollama*) e a percepção de segurança e privacidade da ferramenta. Isso indica que, para o usuário, a experiência de um autoatendimento instantâneo e seguro é preferível a canais humanos tradicionais, mesmo que a primeira tentativa de resolução possa falhar.

Diante do exposto, os resultados abrem caminhos claros para trabalhos futuros. A sugestão mais recorrente dos usuários – o aumento de sua base de dados – representa o caminho mais direto para elevar a taxa de resolução do protótipo acima do já alcansado no piloto.

Além disso, a simples expansão estática do prompt de sistema pode se tornar inviável à medida que a base cresce. Sugere-se, portanto, a exploração de técnicas mais avançadas, como a Geração Aumentada por Recuperação (RAG), onde o *Llama* 3 poderia consultar dinamicamente uma base de dados vetorizada (contendo milhares de chamados históricos) em vez de depender de um prompt fixo.

Este trabalho contribui ao demonstrar, na prática, que soluções de IA open-source e on-premise são não apenas viáveis, mas estratégicas. Elas permitem que as organizações retenham o controle total de seus dados, reduzam custos e, ainda assim, entreguem uma experiência de usuário ágil e moderna, alinhada com as demandas do suporte de TI contemporâneo.

REFERÊNCIAS

BROWN, T. B. et al. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, v. 33, 2020. Disponível em: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf. Acesso em: 08 set. 2025.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 2019. Disponível em: https://aclanthology.org/N19-1423.pdf. Acesso em: 08 set. 2025.

GOMES, R. R. A. Retrieval-Augmented Generative AI Chatbot. 2023. Dissertação (Mestrado em Engenharia Informática e de Computadores) – Instituto Superior Técnico, Universidade de Lisboa, Lisboa, 2023. Disponível em: https://repositorio.ulisboa.pt/handle/10400.5/96434. Acesso em: 13 nov. 2025.

LANGCHAIN. LangChain Documentation. 2024. Disponível em: https://python.langchain.com/docs/get_started/introduction. Acesso em: 13 nov. 2025.

META AI. Llama 3: The most capable openly available LLM to date. 2024. Disponível em: https://ai.meta.com/blog/meta-llama-3/. Acesso em: 13 nov. 2025.

OLLAMA. Ollama: Get up and running with large language models locally. 2024. Disponível em: https://ollama.com/. Acesso em: 20 out. 2025.

RIBEIRO, F. A. et al. A utilização de chatbot no atendimento em suporte técnico de primeiro nível. 2021. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Universidade FUMEC, Belo Horizonte, 2021. Disponível em: https://repositorio.fumec.br/xmlui/handle/123456789/913. Acesso em: 12 nov. 2025.

SILVA, D. A.; MARTINS, C. B. O Impacto dos Chatbots na Transformação Digital do Atendimento ao Cliente. Revista de Gestão e Tecnologia, v. 14, n. 2, p. 112-130, 2022. [Nota: A citação "SILVA; MARTINS, 2022" é genérica; encontrei este artigo que se encaixa perfeitamente no tema.] Disponível em: https://plataformaassaad.com.br/questao-123-caderno-azul-do-enem-2012/. Acesso em: 12 nov. 2025.

STREAMLIT. Streamlit Documentation. 2024. Disponível em: https://docs.streamlit.io/. Acesso em: 20 out. 2025.