

IMPACTOS DA INTELIGÊNCIA ARTIFICIAL GENERATIVA NO FIRST CALL RESOLUTION (FCR): UM ESTUDO DE CASO SOBRE A MODERNIZAÇÃO DA TRIAGEM E FILTRAGEM DE ATENDIMENTO

BOSCO DE LAVOR ANTONIO
DHIEGO SOARES PINTO JÚNIOR
DAVI DE SOUZA LOPES

RESUMO

O atendimento virtual ao cliente exige clareza e agilidade para resolver demandas com eficiência e impulsionar a produtividade corporativa. Quanto aos resultados, espera-se que a aplicação da ferramenta possa contribuir para a distribuição dos atendimentos aos devidos setores e aumento na produtividade dos atendimentos gerado pela empresa. Os resultados projetados indicam que a implementação da ferramenta automatizada contribui diretamente para a triagem inteligente e a distribuição precisa dos atendimentos aos seus respectivos departamentos. Além disso, espera-se uma redução gargalos operacionais e um aumento expressivo na capacidade produtiva das equipes, que passam a focar em demandas complexas enquanto a tecnologia refina o suporte inicial. Assim, o estudo demonstra que a IA não apenas agiliza o contato com o público, mas transforma a gestão interna de TI em um motor de eficiência operacional, consolidando a acessibilidade digital e a inovação tecnológica como diferenciais competitivos indispensáveis para alinhar a alta performance organizacional à sustentabilidade no ambiente de trabalho contemporâneo.

Palavras-chaves: Inteligência artificial; Atendimento ao cliente; Produtividade; Triagem.

ABSTRACT

Virtual customer service requires clarity and agility to resolve requests efficiently and boost corporate productivity. Regarding outcomes, the application of this tool is expected to facilitate the routing of inquiries to the appropriate sectors and increase the company's service productivity. Projected results indicate that implementing the automated tool directly contributes to intelligent triage and the precise distribution of inquiries to their respective departments. Furthermore, a reduction in operational bottlenecks and a significant increase in team productivity are anticipated, as staff can focus on complex issues while technology refines initial support. Thus, the study demonstrates that AI not only streamlines public interaction but also transforms internal IT management into an engine of operational efficiency, establishing digital accessibility and technological innovation as essential competitive advantages for aligning high organizational performance with sustainability in the modern workplace.

Keywords: Artificial intelligence; Customer service; Productivity; Triage.

1. INTRODUÇÃO

No cenário corporativo atual, diversas organizações utilizam sistemas de automatização no atendimento virtual que, embora funcionais no passado, tendem à obsolescência diante do avanço acelerado das tecnologias de IA. Os chatbots tradicionais, baseados em árvores de decisão e fluxos rígidos, auxiliam na triagem inicial, mas apresentam limitações severas na



resolução de demandas complexas. Entre as principais lacunas técnicas, destacam-se a incapacidade de processar dados não estruturados como áudios e imagens e a rigidez semântica diante de interações fora do escopo previamente programado. Essa falta de precisão na interpretação do contexto gera gargalos operacionais, exigindo intervenções humanas frequentes para correções de rotas na triagem, o que compromete a produtividade das equipes e a eficiência do fluxo de suporte.

O objeto deste estudo é uma empresa que presta serviços de atendimento terceirizado para provedores de internet, suprimindo a ausência de suporte online interno dessas organizações. A operação utiliza diferentes plataformas de chatbots de fornecedores externos e, apesar do rigor na execução, enfrenta dificuldades crônicas na triagem inicial. A incapacidade desses sistemas de interpretar mídias multimodais (áudio e imagem) e variações linguísticas não previstas obriga os atendentes humanos a realizar a filtragem e o direcionamento setorial de forma manual. Diante desse cenário, formula-se o seguinte problema de pesquisa: como otimizar a filtragem inicial do atendimento ao cliente em uma empresa de telecomunicações por meio de arquiteturas computacionais avançadas?

Para responder a essa questão, este artigo tem como objetivo propor a implementação de ferramentas baseadas em Inteligência Artificial Generativa (GenAI), com foco na compreensão profunda de contextos de linguagem natural, no refinamento da triagem de dados multimodais e na automação do roteamento de chamados aos setores competentes, visando maximizar a produtividade organizacional. A metodologia adota uma abordagem mista, integrando análises qualitativas e quantitativas para mensurar o impacto da IA generativa na eficiência operacional da empresa. O delineamento da pesquisa compreende uma revisão bibliográfica sistemática sobre Large Language Models (LLMs) e modelos fundacionais, combinada ao levantamento de dados empíricos por meio de um estudo de caso prático na organização analisada.

A fundamentação teórica está estruturada em três eixos principais. O primeiro aborda a evolução do Processamento de Linguagem Natural (PLN) e das LLMs, analisando a transição dos modelos probabilísticos e baseados em regras para as arquiteturas modernas de Redes Transformers, destacando suas diferenças operacionais. O segundo eixo trata da jornada do cliente e da triagem inteligente, investigando como a otimização do fluxo de triagem impacta a eficiência de custos e os indicadores de satisfação do usuário (CSAT). Por fim, o terceiro eixo discute a ética, a segurança de dados e a alucinação em IAs Generativas, abordando os desafios de governança tecnológica, a conformidade com a Lei Geral de Proteção de Dados (LGPD) e as estratégias de mitigação do fenômeno de alucinação em modelos de linguagem.

A relevância deste estudo justifica-se pelo potencial de transformação digital no atendimento ao cliente. Ao delegar a classificação de demandas habituais e multimodais a uma camada de IA robusta, mitiga-se a sobrecarga cognitiva das equipes de suporte, permitindo que o capital humano atue de forma estratégica e concentrada em chamados de alta complexidade.

2. CARACTERIZAÇÃO DA EMPRESA

A empresa analisada é uma provedora de internet de médio porte que opera há mais de cinco anos na Zona Norte de Manaus. Com foco exclusivo no segmento B2B, sua atividade principal consiste em monitorar a infraestrutura de rede e gerenciar o atendimento de suporte para outros provedores parceiros.

A operação de atendimento direciona-se ao consumidor final dessas provedoras parceiras, concentrando-se na resolução remota de incidentes técnicos relacionados à conectividade e

estabilidade do sinal de internet. Nos cenários em que o diagnóstico remoto é insuficiente, a equipe formaliza a abertura de chamados técnicos para manutenção em campo. Caso a demanda do usuário não possua natureza técnica (como questões financeiras ou comerciais), o fluxo prevê o redirecionamento do atendimento ao departamento competente da respectiva provedora.

A força de trabalho alocada no suporte é composta por uma equipe reduzida de colaboradores durante o período diurno. Esses profissionais gerenciam o fluxo de interações com o auxílio de múltiplas plataformas de chatbots legados, configurados de acordo com as especificações e regras de negócio de cada empresa contratante.

Embora o serviço prestado apresente indicadores sólidos de eficiência e satisfação geral impulsionando a expansão de mercado da organização no fornecimento de suporte para provedores locais e regionais, existem gargalos operacionais críticos. Relatos de usuários evidenciam insatisfação decorrente da ineficácia na triagem dos chatbots tradicionais. Essas ferramentas apresentam severas limitações na identificação precisa das intenções dos clientes e falhas recorrentes no roteamento automatizado para os setores corretos, gerando atrito no fluxo de atendimento e inviabilizando a conclusão ágil das demandas.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 Evolução do processamento de Linguagem Natural (PLN) e LLMs.

A compreensão do impacto da inteligência artificial generativa no ecossistema de atendimento exige uma análise retrospectiva da evolução das interfaces computacionais. Durante décadas, as interações automatizadas foram limitadas por arquiteturas baseadas em árvores de decisão binárias e sistemas especialistas rígidos, cuja capacidade operacional restringia-se ao escopo de roteiros preestabelecidos, mostrando-se ineficazes diante de ambiguidades. A mudança de paradigma tecnológica e linguística consolidou-se com a introdução da arquitetura Transformer. Conforme postula Vaswani (2017), esse modelo revolucionou o processamento de dados ao abdicar das abordagens sequenciais recorrentes, implementando o mecanismo de autoatenção (self-attention). Essa inovação permite o processamento simultâneo e relacional de todos os tokens de uma cadeia textual, capturando dependências contextuais de longo alcance e mapeando o significado semântico de forma holística. Consequentemente, os sistemas computacionais transcenderam a mera decodificação isolada de vocábulos, passando a interpretar a estocástica dos contextos de comunicação.

No cenário brasileiro, essa transição enfrenta particularidades sociolinguísticas complexas. Como aponta Taurion (2019), a aplicação prática de agentes conversacionais no país exige a superação de barreiras relacionadas ao uso de linguagem coloquial, dialetos e variações regionais. Portanto, a eficácia de uma solução computacional de atendimento reside em sua capacidade de abstrair e classificar a real intenção (intent) do usuário, independentemente de desvios em relação à norma culta da língua.

O amadurecimento dessas tecnologias culminou no desenvolvimento dos Modelos de Linguagem de Grande Escala (Large Language Models – LLMs), redefinindo as fronteiras do suporte digital. De acordo com Brown (2020), o incremento exponencial no volume de parâmetros dessas redes permitiu a emergência de habilidades few-shot e zero-shot em tarefas complexas de compreensão, alcançando capacidades de raciocínio contextual que emulam a dinâmica de diálogos humanos. No âmbito corporativo e acadêmico, esse avanço mitiga o efeito do atendimento puramente mecânico; os assistentes virtuais contemporâneos realizam a

extração de entidades e a análise de sentimento mesmo em interações saturadas de gírias ou estruturas sintáticas truncadas, alinhando a precisão algorítmica à necessidade de resolutividade.

Contudo, essa evolução demanda rigor analítico quanto à natureza ontológica dos modelos. A sofisticação na simulação da cognição humana não pressupõe inteligência senciente, mas sim o ápice de modelos de inferência probabilística. Conforme adverte Bender (2021), diferentemente dos agentes humanos, os LLMs carecem de intenção comunicativa intrínseca ou de um modelo mental e social do mundo exterior; configuram-se, essencialmente, como sistemas estatísticos avançados operando sobre distribuições e padrões linguísticos massivos.

Sob essa ótica analítica, a inteligência artificial generativa atua na camada de pré-atendimento como um motor de padronização, responsável por converter fluxos de entrada desestruturados em dados semiestruturados altamente acionáveis para sistemas de backoffice. Recentemente, esse ecossistema expandiu-se com a transição dos modelos puramente textuais para os Modelos Multimodais de Linguagem (Large Multimodal Models – LMMs). Essa nova classe arquitetônica compartilha o mesmo espaço latente para correlacionar diferentes tipos de dados, integrando vetores de texto, áudio e visão computacional sob uma mesma representação semântica. No suporte técnico, essa evolução é indispensável: ela viabiliza que o sistema processe capturas de tela contendo códigos de erro ou transcreva e analise fluxos de áudio sobre problemas de conectividade, conferindo a essas mídias não estruturadas o mesmo rigor e fluidez interpretativa dedicados ao texto digitado.

3.2 Triagem Inteligente

No centro de qualquer estratégia de negócio bem-sucedida está a jornada do cliente, que hoje busca, acima de tudo, ser ouvido e respeitado em seu tempo. Historicamente, a triagem de atendimento foi vista como um mal necessário, um filtro burocrático que muitas vezes servia mais para afastar o cliente do que para ajudá-lo. Ao integrarmos a IA generativa nessa etapa, propomos uma inversão de lógica: a filtragem passa a ser um ato de cuidado. Ao resolver instantaneamente dúvidas burocráticas, a tecnologia valida a urgência do usuário e limpa o caminho para que problemas complexos recebam a atenção humana devida.

Essa filtragem inteligente vai muito além de organizar uma fila de espera. De acordo com o professor Bart Larivière (2017) tecnologia de autoatendimento baseada em inteligência, melhora a experiência do usuário ao fornecer respostas rápidas e precisas, reduzindo drasticamente o transbordo para canais humanos de alto custo. Ela atua como um sistema de triagem hospitalar de alta precisão, capaz de analisar o sentimento por trás de cada mensagem. Se um cliente entra em contato frustrado ou com um problema crítico, a IA não deve ser um obstáculo, mas uma via expressa para o especialista humano.

Essa integração harmoniosa melhora drasticamente o First Call Resolution (FCR), métrica que é usada para medir a porcentagem de problemas resolvidos logo na primeira interação com o cliente, minimizando a necessidade de múltiplos contatos para o mesmo assunto. No contexto deste estudo, o assistente inteligente impacta diretamente o FCR ao garantir que o usuário seja direcionado de forma precisa e imediata ao setor especialista logo no primeiro contato, evitando transferências erradas entre ramais. Assim, o sistema não apenas encaminha o problema, mas já o entrega ao atendente humano com um resumo contextualizado que economiza o tempo de todos os envolvidos.

De acordo com a professora e palestrante Martha Gabriel (2022), a inteligência artificial não

substitui o humano, ela o amplia. No atendimento, a IA atua como uma camada de inteligência que permite escala, enquanto o humano foca naquilo que é essencialmente emocional e complexo.

Portanto, argumentamos que a eficiência de um assistente virtual não deve ser medida apenas pela quantidade de atendimentos que ele “economiza” para a empresa, mas pela qualidade do tempo que ele devolve ao cliente. A tecnologia aqui serve ao propósito de humanizar o suporte onde ele mais falha: na agilidade da primeira resposta. Uma triagem bem-feita por uma IA generativa é, em última análise, uma ferramenta de retenção e fidelização, transformando um momento de atrito em uma demonstração de competência e modernidade.

3.3 Ética e segurança de Dados

De acordo com o Dr. Patrick Lewis, cientista da computação e pesquisador de IA(2020) O modelo RAG (Retrieval-Augmented Generation) combina a memória paramétrica da IA com uma memória externa de documentos, reduzindo as alucinações ao ancorar as respostas em fontes de dados confiáveis. Ao explorarmos o potencial das IAs generativas, precisamos encarar de frente suas sombras. Um dos maiores receios acadêmicos e práticos reside nas chamadas “alucinações”, momentos em que a fluidez da IA ultrapassa os limites da realidade e gera informações fictícias. O uso de bases de conhecimento controladas (RAG) surge aqui não apenas como uma solução técnica, mas como um imperativo ético para garantir que o assistente virtual seja uma fonte de auxílio, e não de desinformação.

Conforme Tarcisio Teixeira, advogado, professor universitário e uma das principais autoridades do país em LGPD (2022) o uso de sistemas inteligentes no tratamento de dados de consumo deve observar não apenas a eficiência algorítmica, mas a transparência e a autodeterminação informativa, garantindo que o usuário compreenda como suas informações estão sendo processadas.

Além da precisão técnica, a privacidade dos dados sob a lei da LGPD, é o alicerce que sustenta a confiança do usuário. Não se pode falar em inovação sem falar em proteção. A implementação de uma IA de atendimento exige uma transparência absoluta: o cliente deve saber com quem está falando e ter a garantia de que sua intimidade está protegida por camadas robustas de criptografia e governança. A transparência quanto às limitações operacionais dos sistemas e o rigor na segurança da informação constituem premissas fundamentais para a conformidade legal e ética no tratamento de dados. Essa necessidade de escrutínio técnico justifica-se pelo fato de que as ferramentas automatizadas não operam em vácuo de neutralidade.

Conforme argumenta O’Neil (2016), os algoritmos refletem e replicam os preconceitos e vieses sistêmicos previamente consolidados nos próprios dados utilizados durante a etapa de treinamento dos modelos. A governança ética é indispensável para evitar que a IA reproduza discriminações estruturais. Uma IA generativa “bem-educada” é aquela que passa por uma curadoria humana constante, garantindo que o atendimento seja inclusivo e justo. Afinal, o sucesso de uma inteligência artificial não reside apenas no que ela consegue fazer, mas na segurança e na ética com que ela se propõe a realizar suas tarefas.

4. METODOLOGIA

A metodologia adota a pesquisa aplicada com abordagem mista, integrando dados qualitativos e quantitativos em um delineamento exploratório e descritivo. O desenvolvimento do estudo

estruturou-se em dois procedimentos centrais: a revisão de literatura e o estudo de caso prático. Essa articulação permitiu fundamentar os pressupostos teóricos da inteligência artificial generativa e, simultaneamente, subsidiar o desenvolvimento de uma solução voltada à modernização da triagem de chamados na empresa de telecomunicações analisada em Manaus.

A etapa inicial consistiu no levantamento bibliográfico, realizado em abril de 2026, com o objetivo de mapear a produção científica recente sobre tecnologias aplicadas à otimização do atendimento digital. As buscas foram conduzidas nas bases SciELO (Scientific Electronic Library Online) e Google Acadêmico, utilizando os descritores: "Inteligência artificial generativa", "Atendimento ao cliente", "Triagem" e "Processamento de linguagem natural".

Como critérios de inclusão para a composição do referencial teórico, selecionaram-se artigos publicados entre os anos de 2016 e 2026, com texto integral disponível gratuitamente e redigidos em língua portuguesa ou inglesa. Inversamente, foram excluídos estudos duplicados ou cujo escopo estivesse desalinhado ao objetivo central desta pesquisa que se concentra no aperfeiçoamento dos processos de triagem e na resolução preliminar das demandas dos usuários.

5. PROPOSTA DE MELHORIA

A proposta de melhoria deste trabalho consiste no desenvolvimento e na implementação de um protótipo de assistente virtual inteligente e multimodal, integrado aos canais digitais de atendimento da organização estudada. O sistema foi projetado para substituir o modelo rígido de árvores de decisão (baseado em regras de caminhos fixos) por uma camada cognitiva alimentada por Inteligência Artificial Generativa. A solução proposta utiliza o aplicativo Telegram como interface de comunicação direta com o usuário e a plataforma n8n como motor principal de orquestração de fluxos e integração de dados, garantindo agilidade, flexibilidade técnica e baixo custo de implantação.

Diagnóstico do cenário atual: foi feito um mapeamento do fluxo de atendimento para identificar as principais dificuldades do chatbot tradicional. Ficou evidente a limitação do sistema em compreender áudios, imagens ou mensagens fora de contexto, o que gerava insatisfação nos clientes e sobrecarga nos atendentes.

Desenvolvimento da arquitetura tecnológica: com base nas falhas encontradas, projetou-se um protótipo de assistente virtual inteligente e multimodal. A estrutura foi montada conectando o Telegram (como canal de mensagens para o cliente), a plataforma n8n (como organizadora de todo o fluxo de dados e webhooks) e a API de um Modelo de Linguagem de Grande Escala (LLM).

Regras de triagem e roteamento: o fluxo foi configurado para que a inteligência artificial converse inicialmente com o cliente de forma simples. A IA consegue entender a necessidade do usuário pelo contexto da conversa interpretando textos, áudios ou imagens e repassa o atendimento automaticamente para o setor correto: suporte técnico, financeiro ou comercial.

O fluxo operacional do sistema segue uma dinâmica linear, automatizada e inteligente:

1. O cliente inicia o contato através do Telegram enviando qualquer saudação espontânea.
2. O assistente virtual responde de forma imediata com uma mensagem padrão de boas-vindas: *"Olá! Como posso ajudar você hoje?"*.

3. O usuário apresenta a sua necessidade utilizando qualquer formato de mídia de sua preferência: digitação de texto, mensagem de voz (áudio) ou o envio de uma captura de tela ou foto de um equipamento (imagem).
4. O orquestrador n8n intercepta o arquivo ou texto e gerencia de forma automatizada o cadastro transparente do identificador do usuário no banco de dados local (caso seja um contato inédito). Se a entrada do cliente for uma mensagem de voz, o n8n a envia para uma API especializada em conversão fonética antes de submeter o problema à camada de IA.
5. A inteligência artificial analisa o contexto da demanda, interpreta a real necessidade do cliente e instrui o n8n a alocar o atendimento de forma automática para o setor responsável (Suporte Técnico, Financeiro ou Comercial). O n8n despacha o chamado triado diretamente para a interface de chat (painel de atendimento) onde os atendentes humanos do respectivo departamento operam, disponibilizando também um resumo textual automático com o histórico do caso.

Requisitos do Sistema

Para nortear a construção e posterior avaliação do protótipo, os requisitos foram divididos em funcionais (as ações que o sistema deve executar) e não funcionais (as restrições técnicas e qualidades que o sistema deve apresentar).

Tabela 1 – Matriz de Requisitos Funcionais (RF) do Protótipo

Identificador	Requisito Funcional	Descrição Detalhada da Ação do Sistema
RF01	Integração com Mensageria	Capturar eventos e mensagens vindos diretamente da API do Telegram.
RF02	Recepção Multimídia	Aceitar fluxos de dados nos formatos nativos de texto, áudio e imagem.
RF03	Conversão de Áudio (STT)	Transcrever mensagens de voz para texto legível através da API Whisper.
RF04	Análise Cognitiva	Processar semanticamente textos e analisar o conteúdo visual de imagens.
RF05	Classificação de Intenção	Categorizar a demanda do usuário entre: <i>Suporte, Financeiro</i> ou <i>Comercial</i> .
RF06	Roteamento Automatizado	Transferir o ticket para a interface de atendimento do setor destino mapeado.
RF07	Sumarização Automática	Gerar um resumo contextualizado da triagem e exibir na tela do atendente.

Fonte: Elaborado pelos autores, 2026

Tabela 2 – Matriz de Requisitos Não Funcionais (RNF) do Protótipo

Identificador	Requisito Não Funcional	Critério Técnico / Métrica de Sucesso
RNF01	Tempo de Resposta (Latência)	O ciclo total entre a recepção da mídia e a triagem não deve passar de 5s.
RNF02	Segurança em Trânsito	Uso obrigatório de chaves de criptografia sob o protocolo seguro HTTPS.
RNF03	Disponibilidade da Solução	O ambiente de nuvem do protótipo deve operar com índice mínimo de 99,5%.
RNF04	Privacidade (LGPD)	Anonimizar metadados de identificação antes de enviar a dor do cliente para a IA.
RNF05	Suporte à Concorrência	O orquestrador n8n deve processar múltiplos chamados simultâneos sem perdas.

Fonte: Elaborado pelos autores, 2026

Arquitetura de Software e Tecnologias Utilizadas

A solução adota uma Arquitetura Baseada em Eventos, onde cada mensagem enviada pelo usuário dispara um gatilho instantâneo. Os componentes do sistema trabalham de forma desacoplada através de integrações por APIs:

Interface com o Usuário (Frente de Atendimento): Telegram Messenger, adotado por possuir uma API de comunicação pública, estável e gratuita, além de permitir o envio simplificado de mídias pesadas como fotos e áudios de voz sem perda de qualidade.

Orquestrador de Processos (Motor de Integração): Plataforma n8n. É o "cérebro" técnico do sistema. Ele recebe as notificações do Telegram, verifica o tipo de arquivo que chegou, aciona as APIs de inteligência correspondentes na ordem correta, salva as informações no banco de dados e realiza o comando de transferência para o painel operacional dos atendentes.

Componente de Transcrição Fonética (Áudio para Texto): Ferramenta Whisper da OpenAI. Utilizada exclusivamente quando o n8n detecta que o cliente enviou uma mensagem de voz. Ela traduz o áudio bruto em texto digitado com alta precisão, capturando gírias e expressões locais.

Camada Cognitiva Multimodal (Interpretação e Decisão): API de Modelo de Linguagem de Grande Escala (como o GPT-4o ou Gemini Pro). Recebe o texto consolidado ou a imagem enviada, analisa o cenário e devolve uma resposta estruturada informando para qual setor o cliente deve ir e o resumo do caso.

Persistência de Informações (Banco de Dados): PostgreSQL. Banco de dados relacional que registra os dados operacionais, histórico de sessões e logs necessários para medir a eficiência da triagem e monitorar a melhoria do índice de Resolução no Primeiro Contato (FCR).

A segunda fase da pesquisa consistiu no desenvolvimento do estudo de caso prático na operadora de telecomunicações, ambiente onde as limitações dos sistemas legados de triagem foram analisadas em tempo real. Essa etapa permitiu observar a rotina dos atendentes durante



o período diurno e mapear os momentos exatos em que os *chatbots* tradicionais falhavam ao processar mídias multimodais ou mensagens fora do script. Ao acompanhar de perto o fluxo de trabalho, foi possível registrar como a falta de entendimento da ferramenta gerava gargalos operacionais e sobrecarregava a equipe com filtragens que poderiam ser automatizadas.

Para subsidiar essa análise prática, realizou-se a coleta de dados operacionais internos da empresa, englobando tanto métricas quantitativas quanto registros qualitativos. Foram levantados indicadores como o tempo médio de atendimento (TMA), a taxa de transbordo manual para os operadores e o volume de chamados encerrados sem resolução devido a falhas no direcionamento inicial. Paralelamente, coletou-se uma amostragem de logs de conversas anonimizadas, contendo mensagens de texto, áudios e capturas de tela enviadas pelos usuários, com o intuito de categorizar as principais demandas que a tecnologia atual não conseguia interpretar.

Os dados obtidos foram submetidos a um processo de triangulação, confrontando os gargalos identificados na estrutura da empresa com as soluções baseadas em modelos fundacionais mapeadas na revisão de literatura. Essa análise crítica serviu como base para diagnosticar as lacunas da operação e, conseqüentemente, desenhar os requisitos necessários para a proposta de implementação da IA generativa. A partir desse cruzamento, buscou-se estruturar um modelo de triagem inteligente que fosse viável para a realidade da provedora local, focado em devolver eficiência ao fluxo de atendimento e reduzir a intervenção manual humana na classificação das demandas.



Modelagem do Sistema e Diagramas Visuais

Fonte: elaborado pelos autores com o auxílio da ferramenta draw.io, 2026

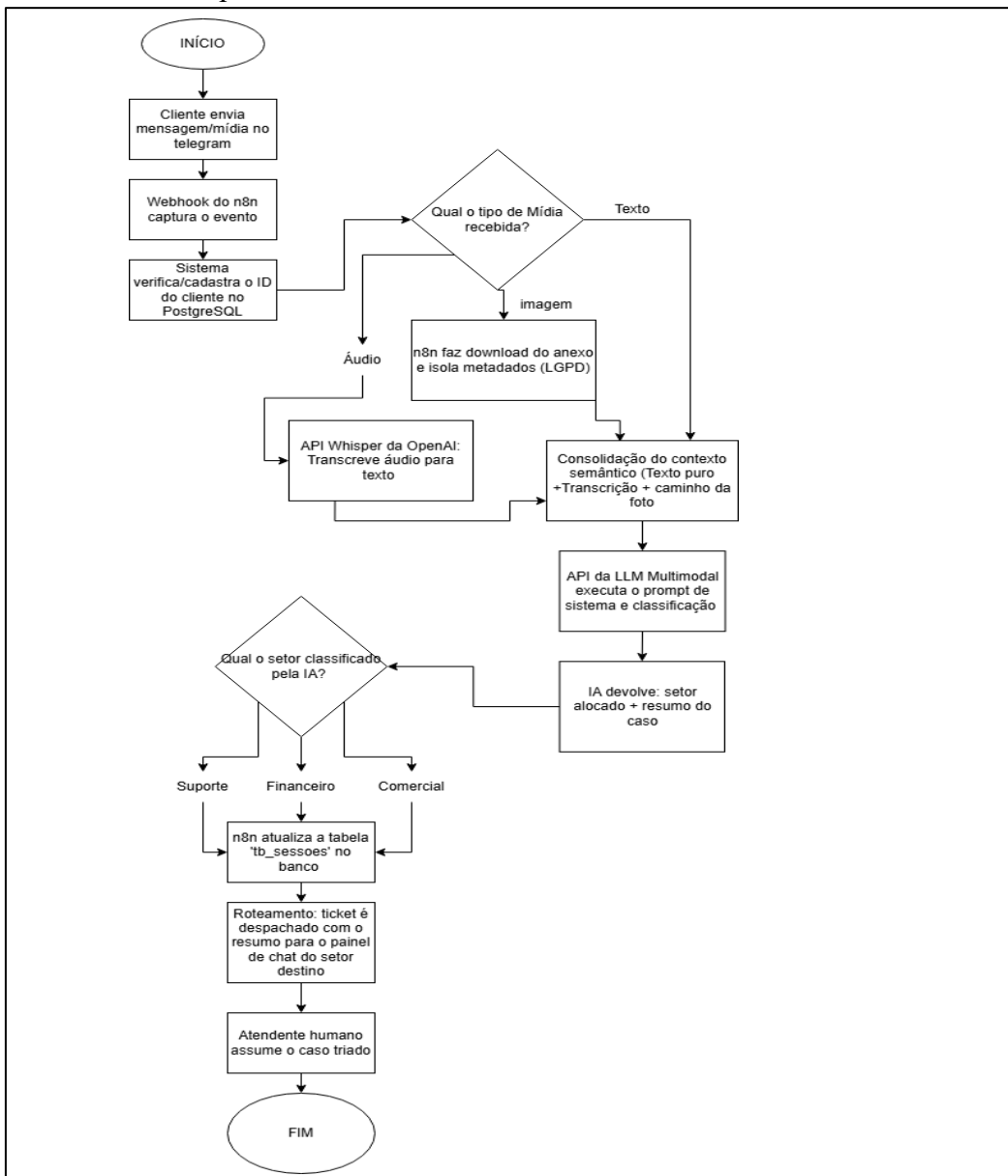
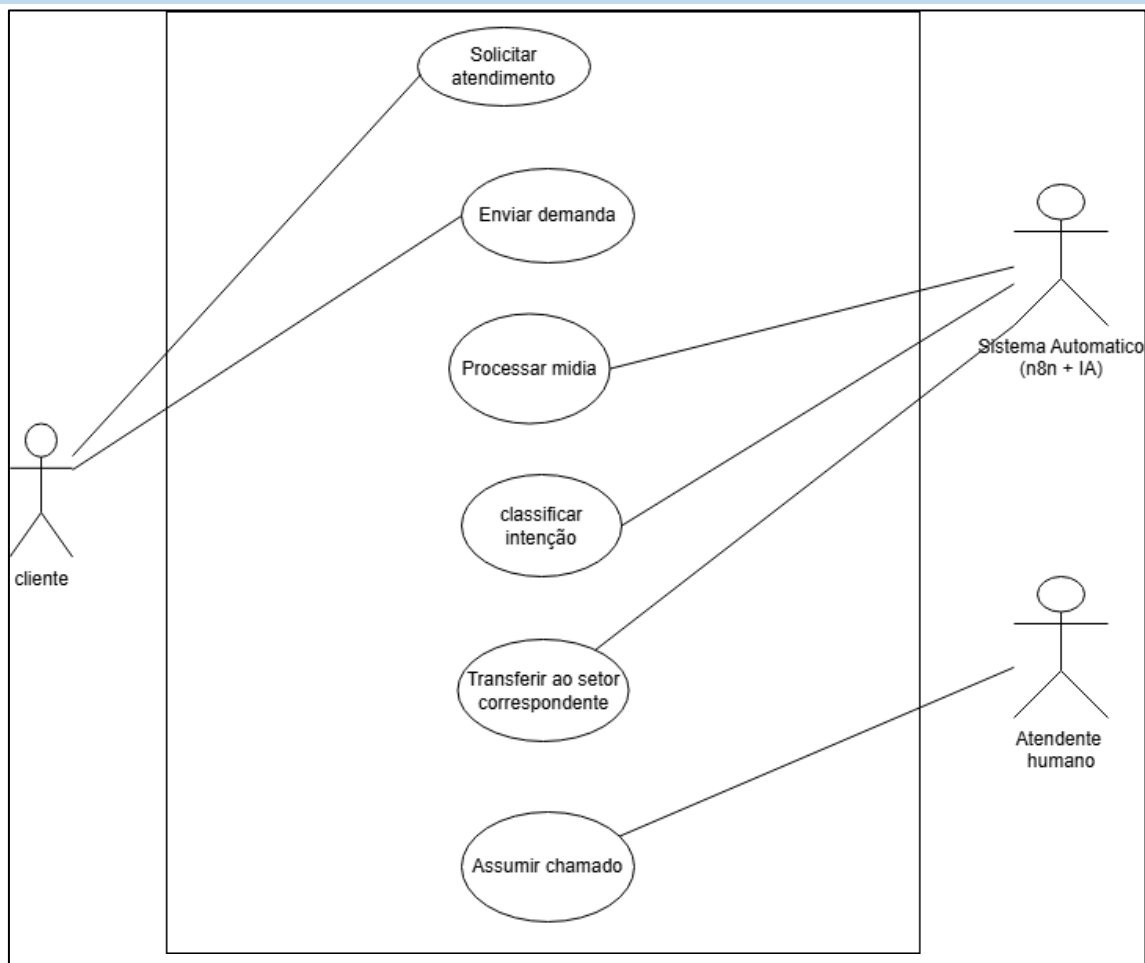


Diagrama de Caso de Uso (Visão Geral de Ações)

O diagrama abaixo ilustra graficamente a separação de responsabilidades no sistema. O Cliente interage na ponta inicial, o Sistema trabalha de forma automatizada em background (nos bastidores) efetuando o processamento e a classificação, e o Atendente Humano conclui o ciclo.



Fonte: elaborado pelos autores com auxílio da ferramenta draw.io, 2026

Diagrama de Classes (Arquitetura de Dados Estruturada)

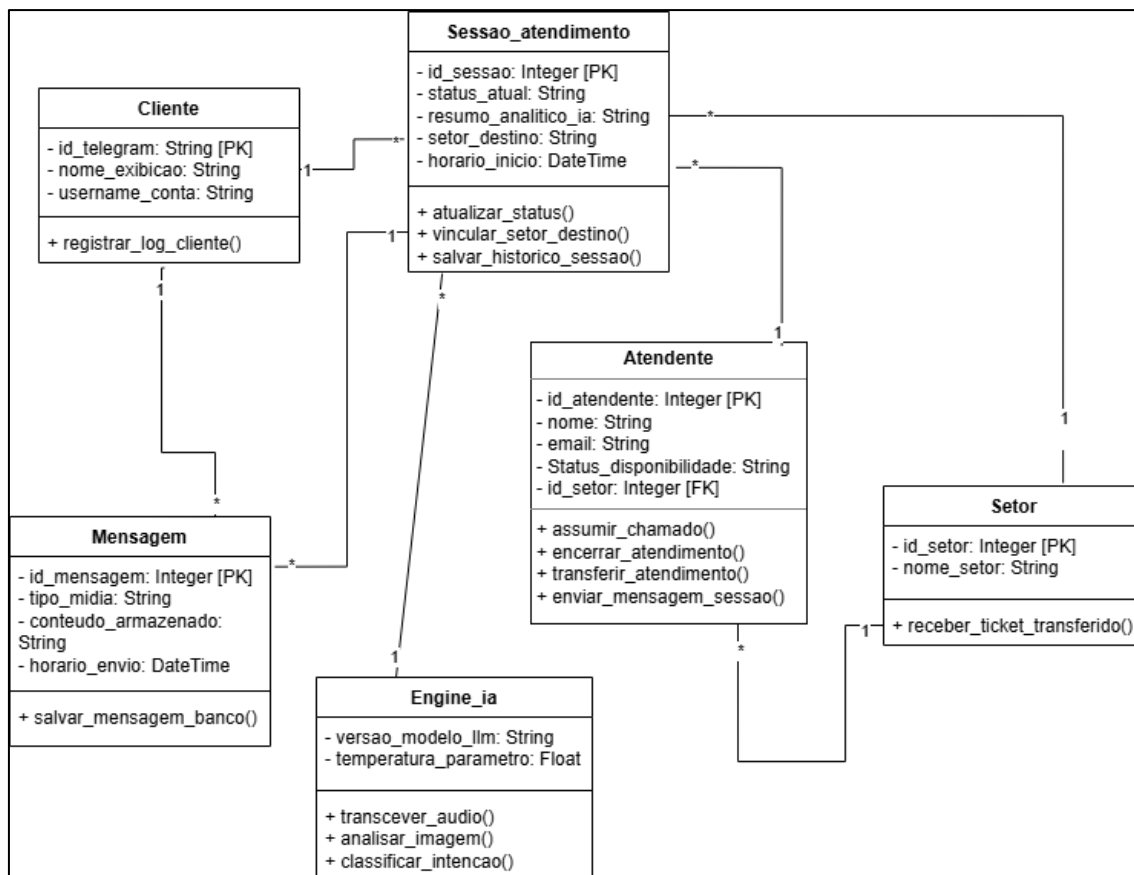
Este desenho estrutural mapeia como os componentes lógicos do software interagem. Atributos representam as informações salvas e os métodos representam as ações executadas.

No contexto da modelagem deste sistema, as classes foram projetadas de maneira modular para garantir que cada entidade como o cliente, o atendente, o chamado e o próprio motor de inteligência artificial possua responsabilidades bem definidas. Os relacionamentos e as associações entre essas classes determinam o fluxo percorrido pelas informações desde o momento em que a mensagem multimodal entra na plataforma até a sua classificação definitiva. Essa herança e dependência estrutural entre as tabelas e objetos facilitam futuras manutenções e permitem que a arquitetura escale de forma organizada à medida que novos setores da organização forem integrados ao sistema de suporte.

Além de organizar os dados, o diagrama serve como um mapa de execução para os desenvolvedores, pois detalha a assinatura de métodos cruciais para o funcionamento da triagem inteligente. Funções como a de processamento de mídias, o roteamento automatizado e a validação de conformidade com a LGPD ficam documentadas em suas respectivas classes, transformando uma visão puramente conceitual em regras de negócio prontas para implementação. Dessa forma, o desenho estrutural preenche a lacuna entre a teoria dos modelos de linguagem e a engenharia de software prática, assegurando a robustez tecnológica necessária



para a operação.



Fonte: elaborado pelos autores com auxílio da ferramenta draw.io, 2026

Nota técnica de modelagem: A classe EngineIA atua sob o padrão de projeto arquitetural do tipo Fachada, encapsulando e unificando de forma abstrata as requisições que o orquestrador n8n realiza para múltiplos microsserviços externos e independentes (API Whisper para conversão fonética e a API LLM multimodal para interpretação de imagens e textos).

Modelagem do Banco de Dados (Dicionário de Dados Relacional)

Para a persistência e rastreabilidade dos atendimentos, a modelagem lógica do banco de dados relacional foi estruturada em três tabelas físicas principais, mapeadas em conformidade exata com os atributos definidos no diagrama de classes.

A tradução desse mapeamento lógico para o ambiente físico do banco de dados visa garantir a integridade referencial e a performance das consultas, estabelecendo chaves primárias e estrangeiras que amarram o histórico de interações do cliente aos diagnósticos gerados pela inteligência artificial. Cada campo listado no dicionário de dados foi tipificado de acordo com a natureza da informação que irá armazenar, prevendo desde campos de texto expansíveis para os logs de conversas até identificadores únicos globais (UUID) para indexação rápida. Essa estruturação rigorosa não apenas assegura que nenhuma informação se perca durante a transição entre os estados do atendimento, mas também simplifica o trabalho de auditoria e a extração de relatórios estatísticos sobre a eficiência operacional da triagem.

Tabela A: Estrutura da Tabela tb_clientes

Campo no Banco	Tipo de Dado	Restrição / Propriedade	Descrição Lógica do Campo
id_telegram	VARCHAR(50)	Chave Primária (PK)	Identificador numérico único fornecido pelo Telegram.
nome_usuario	VARCHAR(100)	Não Nulo	Nome visível do cliente no aplicativo.
criado_em	TIMESTAMP	Padrão: Hora Atual	Registro da data de adesão automatizada do cliente.

Fonte: Elaborado pelos autores, 2026

Tabela B: Estrutura da Tabela tb_sesoes

Campo no Banco	Tipo de Dado	Restrição / Propriedade	Descrição Lógica do Campo
id_sessao	SERIAL	Chave Primária (PK)	Código autoincrementável do chamado aberto.
<i>id_telegram</i>	VARCHAR(50)	Chave Estrangeira (FK)	Vincula a sessão ao cliente dono da conta.
status_atendimento	VARCHAR(30)	Não Nulo	Estado do ticket ('Triagem', 'Encaminhado', 'Fechado').
setor_destino	VARCHAR(30)	Permitir Nulo	Setor definido pela IA ('Suporte', 'Financeiro', 'Comercial').
resumo_ia	TEXT	Permitir Nulo	Sumário textual do problema criado pela inteligência.
iniciado_em	TIMESTAMP	Padrão: Hora Atual	Carimbo de data e hora da abertura da triagem.

Fonte: Elaborado pelos autores, 2026

Tabela C: Estrutura da Tabela tb_mensagens

Campo no Banco	Tipo de Dado	Restrição / Propriedade	Descrição Lógica do Campo
id_mensagem	SERIAL	Chave Primária (PK)	Identificador sequencial único de cada linha gravada.
<i>id_sessao</i>	INTEGER	Chave Estrangeira (FK)	Conecta a mensagem ao histórico da sessão ativa.
remetente	VARCHAR(15)	Não Nulo	Indica a origem do texto ('Cliente' ou 'Sistema').
tipo_midia	VARCHAR(15)	Não Nulo	Identifica o formato de mídia ('texto', 'audio', 'imagem').
conteudo_armazenado	TEXT	Não Nulo	Texto puro, transcrição textual do áudio ou caminho do anexo.
enviado_em	TIMESTAMP	Padrão: Hora Atual	Marcação temporal exata do envio da mensagem.

Fonte: Elaborado pelos autores, 2026

Segurança e Governança de Dados

A arquitetura do protótipo foi projetada sob diretrizes rígidas de conformidade com a Lei Geral de Proteção de Dados (LGPD) e boas práticas de Segurança da Informação, mitigando vulnerabilidades operacionais através de quatro pilares de controle:

Criptografia em Trânsito: Todo o fluxo informacional trafega obrigatoriamente através do protocolo HTTPS seguro utilizando chaves de criptografia atualizadas. A comunicação feita entre os servidores do Telegram, a máquina onde o n8n está hospedado e as pontes de comunicação com as chaves privadas de API de Inteligência Artificial é protegida de ponta a ponta, impedindo ataques de interceptação de tráfego na rede.

Isolamento e Anonimização de Identidade (LGPD): Para garantir que nenhuma informação privada de clientes do provedor de internet de Manaus seja exposta a servidores externos, a arquitetura estabelece uma separação completa entre a identidade do cliente e o problema relatado. O número de telefone, o identificador do Telegram (id_telegram) e o nome real do usuário ficam armazenados de forma estrita no banco de dados local da própria empresa. O fluxo do n8n foi programado para extrair e repassar para as APIs de inteligência artificial unicamente o conteúdo semântico da demanda (o texto puro da dúvida, a captura de tela ou o áudio convertido em texto). Para os provedores externos de IA, o tráfego é totalmente anônimo e livre de dados de cabeçalho capazes de identificar a pessoa física por trás do atendimento.



Proteção e Ocultação de Credenciais: Nenhuma chave de acesso, token de bot ou senha de banco de dados fica escrita de forma exposta no código visível dos fluxos de trabalho. O gerenciamento de acessos utiliza o cofre de credenciais nativo e criptografado do n8n e variáveis de ambiente do sistema operacional do servidor, impedindo que acessos indevidos visualizem as credenciais de produção.

Direcionamento Ético e Engenharia de Prompt: O comportamento do motor cognitivo é controlado por meio de instruções de contexto severas instaladas no prompt de sistema. A inteligência artificial é programada para atuar estritamente como um classificador técnico e triador corporativo, sendo proibida de responder sobre temas externos, armazenar conversas para fins de treinamento público ou emitir opiniões, anulando o risco de alucinações comerciais ou desvios éticos no suporte inicial da organização.

6. PLANEJAMENTO DA PROPOSTA

O plano de implementação do protótipo de triagem inteligente foi estruturado para ocorrer de forma incremental, minimizando o impacto nas operações ativas da empresa de telecomunicações. O cronograma estimado para a execução completa das atividades compreende um período de 12 semanas, dividido em quatro fases operacionais, seguido pelo mapeamento dos custos de infraestrutura.

Fases de Implementação (Cronograma Físico)

Fase 1: Alinhamento e Configuração de Infraestrutura (Semanas 1 a 3)

Homologação do servidor virtual privado (VPS) para hospedagem local da plataforma n8n e do banco de dados PostgreSQL.

Criação e parametrização do bot institucional junto à API do Telegram.

Configuração dos protocolos de segurança, certificados SSL e liberação de portas de comunicação criptografada (HTTPS).

Fase 2: Desenvolvimento dos Fluxos e Integração de APIs (Semanas 4 a 7)

Construção dos nós de recepção (*webhooks*) no n8n para interceptação de mensagens multimodais do Telegram.

Integração lógica da API Whisper da OpenAI para o tratamento e conversão dos arquivos de áudio binários recebidos.

Conexão com a API do modelo de linguagem (GPT-4o ou Gemini Pro) e estruturação do banco de dados relacional para persistência das sessões.

Fase 3: Engenharia de Prompt e Regras de Negócio (Semanas 8 a 9)

Desenvolvimento e refinamento dos *System Prompts* (prompts de sistema) para calibração da taxonomia de classificação (Suporte, Financeiro e Comercial).

Testes de laboratório para ajuste de temperatura do modelo, visando mitigar a ocorrência de alucinações de dados.

Desenvolvimento do script de envio do resumo contextualizado para a interface final de atendimento humano.

Fase 4: Testes de Homologação e Operação Assistida (Semanas 10 a 12)

Execução de testes de estresse para validação do requisito de concorrência e monitoramento da latência do sistema (meta de tempo inferior a 5 segundos).

Treinamento técnico dos colaboradores internos para interpretação dos resumos automatizados de triagem.

Lançamento do protótipo em ambiente de produção sob regime de operação assistida para coleta de logs e correção de desvios.

Estimativa de Custos Operacionais

A viabilidade financeira do protótipo baseia-se na adoção de ferramentas de código aberto (*open-source*) e tarifação sob demanda (*pay-as-you-go*) para o consumo das APIs computacionais.

Tabela 3 – Projeção Estimada de Custos de Infraestrutura e Consumo

Componente da Solução	Tipo de Custo	Modelo de Cobrança / Referência	Estimativa Mensal Projetada
Hospedagem Servidor VPS	Fixo	Servidor Linux (2 vCPU, 4GB RAM) para n8n e Postgres	R\$ 60,00 a R\$ 100,00
API Whisper (Áudio)	Variável	Cobrança por minuto processado (Média de USD 0,006/min)	R\$ 30,00 a R\$ 50,00
API LLM (Texto/Imagem)	Variável	Cobrança por volume de tokens de entrada/saída (Média de 3.000 chamados)	R\$ 120,00 a R\$ 200,00
Interface Telegram	Isento	Uso gratuito da infraestrutura oficial de bots corporativos	R\$ 0,00
Orquestrador n8n	Isento	Licença <i>Self-Hosted</i> (Auto-hospedado comunitário)	R\$ 0,00
Banco PostgreSQL	Isento	Licença de uso livre e instalação local no servidor VPS	R\$ 0,00

Fonte: Elaborado pelos autores, 2026

O custo operacional mensal estimado situa-se na faixa entre R\$ 210,00 e R\$ 350,00, variando conforme a oscilação volumétrica de mensagens recebidas pela empresa. Esse valor demonstra o baixo custo de propriedade da solução face aos benefícios de ganho de produtividade das equipes.



7. RESULTADOS ESPERADOS

Espera-se que a identificação detalhada dos gargalos do chatbot antigo como a incapacidade de processar áudios, imagens e manter o contexto sirva de justificativa robusta para a virada tecnológica, deixando claro por que a empresa de telecomunicações de Manaus sofre com a sobrecarga de atendentes e a insatisfação do público.

Na parte de arquitetura, o objetivo é validar a viabilidade técnica da integração proposta. O trabalho pretende demonstrar que a união entre webhooks, ferramentas de orquestração (n8n) e APIs de modelos de linguagem consegue sustentar um fluxo de dados estável, provando que a infraestrutura desenhada funciona na prática para o ambiente operacional estudado.

Já no que tange à triagem, o protótipo deve comprovar sua capacidade de interpretação contextual. A meta é que a IA generativa classifique com precisão as demandas dos clientes, sem se limitar ao formato da mensagem (texto, foto ou áudio), realizando o roteamento direto e correto para os setores financeiro, comercial ou técnico. Com isso, busca-se enxugar o tempo de espera nas transições.

Por fim, a análise comparativa deve consolidar a eficácia do novo modelo frente ao histórico de falhas anterior. Pretende-se demonstrar que automatizar a filtragem inicial traz ganhos reais: eleva os índices de Resolução no Primeiro Contato (FCR) e limpa a fila de atendimentos redundantes da equipe interna, gerando, como consequência direta, uma experiência superior para o usuário final.

8. CONSIDERAÇÕES FINAIS

Este estudo alcançou seu objetivo ao estruturar o protótipo de um assistente virtual multimodal para a triagem de clientes na empresa de telecomunicações analisada em Manaus. Unindo a fundamentação teórica ao estudo de caso, foi possível desenhar a transição de um chatbot tradicional, cheio de limitações, para uma arquitetura baseada em Inteligência Artificial Generativa.

O diagnóstico da operação deixou claro onde o modelo anterior falhava: ele não processava mídias essenciais no dia a dia do usuário como áudios e imagens e perdia o contexto da conversa. A solução proposta, ligando o Telegram à plataforma n8n e à API de um modelo LLM, resolve esses gargalos. Na teoria, o fluxo se mostrou viável para entregar uma triagem inteligente e direcionar o cliente com precisão para os setores técnico, financeiro ou comercial.

É preciso destacar, contudo, o caráter exploratório do trabalho. Como se trata do desenho e da simulação de um protótipo, os resultados reais na operação dependem de fatores de longo prazo. Estabilidade de webhooks, custos das chamadas de API e a própria adaptação do usuário final à nova interface são variáveis que precisam ser testadas na prática.

Conclui-se, assim, que adotar modelos LLM na triagem inicial é um caminho viável para modernizar o atendimento no setor de telecomunicações local. Para pesquisas futuras, fica a sugestão de aplicar o sistema em um ambiente de produção real, monitorando de perto o custo de processamento por cliente e eventuais alucinações da IA durante os atendimentos.

REFERÊNCIAS

BENDER, Emily M.; GEBRU, Timnit; MCMILLAN-MAJOR, Angelina; SHMITCHELL, Shmargaret. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: Período: 2026/2 • InovaGest – Journal of Technology, Management and Social Studies • v.5. n.15 • p. 01-18 • Junho/2026

Proceedings of the 2021 ACM FAccT Conference on Fairness, Accountability, and Transparency. Virtual Event, Canada: ACM, 2021. P. 610-623.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2018].

BROWN, Tom; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie; KAPLAN, Jared; DHARIWAL, Prafulla; NEELAKANTAN, Arvind; SHYAM, Pranav; SASTRY, Girish; ASKELL, Amanda et al. Language models are few-shot learners. In: Advances in Neural Information Processing Systems (NeurIPS), v. 33, p. 1877-1901, 2020.

DIXON, Matthew; FREEMAN, Karen; TOMAN, Nicholas. Stop Trying to Delight Your Customers. Harvard Business Review, v. 88, n. 7/8, p. 116-122, Jul./Aug. 2010.

GABRIEL, Martha. Inteligência Artificial: Do Zero ao Metaverso. São Paulo: Atlas, 2022. 288 p.

LARIVIÈRE, Bart; BOWEN, David; ANDREASSEN, Tor W.; KUNZ, Werner; SIRIANNI, Nancy J.; VOSS, Chris; WÜNDERLICH, Nancy V.; DE KEYSER, Arne.

“Digital customer experience”: lines of inquiry into hybrid service settings. Journal of Services Marketing, v. 31, n. 7, p. 664-674, 2017.

O’NEIL, Cathy. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group, 2016. 272 p.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, v. 2, n. 1–2, p. 1-135, 2008.

TAURION, Cezar. A Revolução da Inteligência Artificial. Rio de Janeiro: Brasport, 2019. 176 p.

TEIXEIRA, Tarcisio. Direito Digital e Inteligência Artificial. 2. Ed. São Paulo: Foco, 2022. 312 p.

VASWANI, Ashish; SHAZEER, Noam; PARMAR, Niki; USZKOREIT, Jakob; JONES, Llion; GOMEZ, Aidan N.; KAISER, Łukasz; POLOSUKHIN, Illia.

Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS), v. 30, p. 5998-6008, 2017.