

INSTITUTO FEDERAL DE MINAS GERAIS - *CAMPUS* OURO PRETO
CURSO ESPECIALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL

Alexandre Fortes Santana

**APLICAÇÃO DE MODELOS DE LINGUAGEM DE GRANDE ESCALA
(LLM) NA AUTOMAÇÃO COGNITIVA: UMA PROVA DE CONCEITO
EM ATENDIMENTO AO CLIENTE**

Ouro Preto

2024

ALEXANDRE FORTES SANTANA

**APLICAÇÃO DE MODELOS DE LINGUAGEM DE GRANDE ESCALA
(LLM) NA AUTOMAÇÃO COGNITIVA: UMA PROVA DE CONCEITO
EM ATENDIMENTO AO CLIENTE**

Trabalho de conclusão do curso de Especialização em Inteligência Artificial do Instituto Federal de Minas Gerais - *Campus* Ouro Preto para a obtenção do certificado de conclusão.

Orientador: Prof. Frederico Gadelha Guimarães (UFMG)

Ouro Preto
2024

S232a

Santana, Alexandre Fortes.

Aplicação de modelos de linguagem de grande escala (LLM) na automação cognitiva [manuscrito] : uma prova de conceito em atendimento ao cliente / Alexandre Fortes Santana. – 2024.
59 f. : il.

Orientador: Frederico Gadelha Guimarães.

Trabalho de Conclusão de Curso (especialização) – Instituto Federal de Minas Gerais. *Campus* Ouro Preto, 2024.

1. Inteligência artificial. 2. Engenharia de Prompt. 3. Automação Cognitiva. I. Guimarães, Frederico Gadelha. II. Instituto Federal de Minas Gerais. *Campus* Ouro Preto. III. Título.

CDU: 004.8

Catálogo: Kelly Cristiane Santos Moraes - CRB-6/3217



MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS
Campus Ouro Preto - Diretoria de Pesquisa, Inovação e Pós-Graduação
Coordenação do Curso de Pós-graduação em Inteligência Artificial
Rua Pandiá Calógeras, 898 - Bairro Bauxita - CEP 35400-000 - Ouro Preto - MG
- www.ifmg.edu.br

FOLHA DE APROVAÇÃO

ALEXANDRE FORTES SANTANA

APLICAÇÃO DE MODELOS DE LINGUAGEM DE GRANDE ESCALA (LLM) NA AUTOMAÇÃO COGNITIVA: UMA PROVA DE CONCEITO EM ATENDIMENTO AO CLIENTE

Trabalho de Conclusão de Curso apresentado ao curso de ESPECIALIZAÇÃO EM INTELIGÊNCIA ARTIFICIAL, ofertado pelo Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais - *Campus* Ouro Preto, como parte dos requisitos para a obtenção do título de ESPECIALISTA EM INTELIGÊNCIA ARTIFICIAL.

Aprovado em 13 de setembro de 2024, pela Banca Examinadora:

Prof. Dr. Frederico Gadelha Guimarães - UFMG - Colaborador externo - Orientador

Profa. Dra. Andrea Gomes Campos - DECOM - UFOP

Profa. Dra. Silvia Grasiella Moreira Almeida - IFMG - Campus Ouro Preto

Ouro Preto, 03 de outubro de 2024.

AGRADECIMENTOS

Agradeço, em primeiro lugar, à minha esposa, Débora Mendes, por me lembrar constantemente que escrever um trabalho acadêmico não é tão chato ou difícil quanto parece, oferecendo-me apoio e incentivo ao longo deste processo.

Meu sincero agradecimento também ao Tiago Borges, Tech Lead na Efí Bank, e ao Mateus Souza, Engenheiro de Software na Efí Bank, pela atuação fundamental no desenvolvimento da versão final deste projeto.

Agradeço ao professor Frederico Gadelha Guimarães, da UFMG, e à professora Sílvia Grasiella Moreira Almeida, do IFMG - Campus Ouro Preto, pela valiosa orientação e suporte.

Agradeço ao Diego Alberto e ao meu time na Efí Bank. A colaboração e o conhecimento compartilhado por vocês são essenciais para transformar ideias em soluções práticas e inovadoras.

Agradeço ao Evanil Paula e ao Leôncio Nogueira pelo constante apoio à inovação contínua na Efí Bank.

Por fim, expresso minha gratidão a todos os professores da Especialização em Inteligência Artificial do IFMG - Campus Ouro Preto, que me proporcionaram uma formação sólida e essencial para a realização deste projeto.

RESUMO

Este trabalho apresenta a aplicação de modelos de linguagem de grande escala (LLMs), especificamente o GPT-4, na automação da avaliação de atendimentos ao cliente, validando sua eficácia por meio de uma prova de conceito (PoC) implementada em um ambiente real na Efi Bank. A hipótese inicial propôs que o LLM poderia reduzir a necessidade de intervenção humana direta na classificação e análise de chamados, permitindo que ajustes fossem realizados por profissionais não especializados em TI, como líderes de atendimento ou analistas de qualidade. Os resultados corroboram essa hipótese, destacando a flexibilidade dos LLMs em se adaptarem a novas demandas com simples ajustes de prompts. Além disso, o projeto demonstrou que a automação pode realizar milhares de análises com um custo operacional competitivo, liberando os profissionais para focarem em atividades de maior valor agregado. Entre os principais desafios enfrentados, destacam-se a sensibilidade dos LLMs ao formato dos prompts e o fenômeno das “alucinações”, em que o modelo pode gerar dados inexistentes ao tentar responder a uma solicitação inadequada. Para mitigar esses desafios, foram implementadas estratégias de refinamento dos prompts e normalização de dados quantitativos. Os resultados obtidos demonstram que, além de serem economicamente viáveis, os LLMs podem superar as limitações de análises humanas tradicionais em termos de escala e consistência. O projeto também identificou oportunidades de aplicação dessa tecnologia em outras áreas da organização, como na detecção de fraudes e na automação de processos comerciais, ampliando o impacto estratégico da automação cognitiva dentro da empresa.

Palavras-chave: Inteligência artificial generativa. Engenharia de prompt. Modelos de linguagem de grande escala. Automação cognitiva. ChatGPT. Atendimento ao cliente.

ABSTRACT

This work presents the application of large-scale language models (LLMs), specifically GPT-4, in automating customer service evaluation, validating its effectiveness through a proof of concept (PoC) implemented in a real environment at Efi Bank. The initial hypothesis proposed that the LLM could reduce the need for direct human intervention in the classification and analysis of service requests, allowing adjustments to be made by non-IT professionals, such as customer service leaders or quality analysts. The results support this hypothesis, highlighting the flexibility of LLMs in adapting to new demands with simple prompt adjustments. Furthermore, the project demonstrated that automation can perform thousands of analyses at a competitive operational cost, freeing professionals to focus on higher-value activities. Among the main challenges encountered were the sensitivity of LLMs to prompt formatting and the phenomenon of “hallucinations,” where the model may generate non-existent data when attempting to respond to an inappropriate request. To mitigate these challenges, strategies for refining prompts and normalizing quantitative data were implemented. The results obtained demonstrate that, in addition to being economically viable, LLMs can overcome the limitations of traditional human analyses in terms of scale and consistency. The project also identified opportunities to apply this technology in other areas of the organization, such as fraud detection and the automation of commercial processes, expanding the strategic impact of cognitive automation within the company.

Keywords: Generative artificial intelligence. Prompt engineering. GPT-4. Large-scale language models. Cognitive automation. ChatGPT. Customer service.

LISTA DE ILUSTRAÇÕES

Figura 1 – Site Efi Bank. Disponível em https://sejaefi.com.br	15
Figura 2 – Resultado do prompt simples. Crédito: Andrew Best ³⁴	23
Figura 3 – Resultado do prompt mais detalhado. Crédito: Andrew Best ⁴⁴	23
Figura 4 – Fluxograma do processo de atendimento	31
Figura 5 – Visão geral do dashboard com análise quantitativa de tags, produtos, motivos de contato e filtros específicos, permitindo o monitoramento detalhado do desempenho dos atendimentos e a identificação de padrões relevantes.	49
Figura 6 – Distribuição das tags mais utilizadas em atendimentos, destacando a frequência com que cada tag foi aplicada, com “Tag 1” sendo a mais comum.	50
Figura 7 – Quantidade normalizada de menções a produtos durante os atendimentos, com “Produto 1” sendo o mais referenciado em relação aos demais.	50
Figura 8 – Motivos dos contatos classificados em problemas, solicitações, dúvidas e reclamações, com “Problemas” sendo o motivo mais frequente.	51
Figura 9 – Evolução temporal da quantidade normalizada de dois produtos anonimizados, analisando como a menção a esses produtos variou ao longo do tempo.	51
Figura 10 – Evolução temporal do uso de duas tags anonimizadas, destacando as variações na frequência dessas tags ao longo do tempo.	51

LISTA DE ABREVIATURAS E SIGLAS

GPT	Generative Pre-trained Transformer
GTC	GPU Technology Conference
GPU	Graphics Processing Unit
IA	Inteligência Artificial
LLM	Large Language Models
PLN	Processamento de Linguagem Natural
RLHF	Reinforcement Learning with Human Feedback
GRE	Graduate Record Examinations
IoT	Internet das Coisas
CoT	Chain-of-Thought
CRM	Customer Relationship Management
MMLU	Massive Multitask Language Understanding
STEM	Science, Technology, Engineering, and Mathematics
POC	Proof of Concept (Prova de Conceito)

SUMÁRIO

1	INTRODUÇÃO	11
1.1	UMA PROJEÇÃO DO MUNDO	11
1.2	ENTENDIMENTO DE CONTEXTOS	12
1.2.1	Engenharia de prompt	13
1.3	HIPÓTESES	14
1.4	OBJETIVOS	14
1.4.1	Objetivo geral	14
1.4.2	Objetivos específicos	14
1.5	EFÍ BANK	15
2	REVISÃO DA LITERATURA	16
2.1	INTELIGÊNCIA HUMANA VERSUS INTELIGÊNCIA ARTIFICIAL	16
2.2	ASPECTOS COGNITIVOS E ARQUITETURAIS DO GPT-4 NA GE- NERALIZAÇÃO DA INTELIGÊNCIA ARTIFICIAL	16
2.3	VANTAGENS DOS LLMS	17
2.3.1	Escalabilidade	17
2.3.2	Adaptabilidade	18
2.3.3	Geração de conteúdo	18
2.3.4	Multimodalidade	18
2.4	LIMITAÇÕES E DESAFIOS DOS LLMS	18
2.4.1	Viés e preconceito	18
2.4.2	Custo computacional	18
2.4.3	Interpretação e explicabilidade	19
2.4.4	Segurança e uso indevido	19
2.5	PERSPECTIVAS FUTURAS E TENDÊNCIAS	19
2.5.1	Avanços na arquitetura	19
2.5.2	Modelos mais sustentáveis	19
2.5.3	Regulamentação e governança	19
2.5.4	Adição de uma camada de raciocínio nos LLMs	19
2.6	ENGENHARIA DE PROMPT	20
2.6.1	Especificação de instruções claras	20
2.6.2	Utilização de texto de referência	21
2.6.3	Divisão de tarefas complexas em subtarefas simples	22
2.6.4	Tempo para processamento da resposta	22
2.6.5	Limitações da adição excessiva de detalhes em prompts	22

2.6.6	Desafios Éticos e de Propriedade Intelectual na Geração de Conteúdos por IA	24
3	METODOLOGIA	25
3.1	TIPO DE PESQUISA	25
3.2	MÉTODOS DE AVALIAÇÃO E VALIDAÇÃO	25
3.3	FERRAMENTAS E TECNOLOGIAS UTILIZADAS	25
3.4	PROVA DE CONCEITO (POC)	25
3.4.1	Critérios de avaliação de sucesso	25
3.4.2	Coleta e preparação de dados	25
3.4.3	Implementação dos testes	28
3.4.4	Análise dos resultados	28
3.5	IMPLEMENTAÇÃO DO PROJETO	29
3.5.1	Escopo de canais para a avaliação	29
3.5.2	Visão geral do processo de atendimento omnichannel e con- sumo da API	31
3.5.3	Prompt construído	32
3.5.4	Discussão sobre o prompt	35
3.5.4.1	Clareza e especificação do contexto	35
3.5.4.2	Estrutura e organização	35
3.5.4.3	Especificação do formato de saída	35
3.5.4.4	Normalização de respostas e limitação do escopo	36
3.5.4.5	Verificação de respostas	36
3.5.4.6	Campos qualitativos	36
3.5.4.7	Assunto	36
3.5.4.8	Sugestão	37
3.5.4.9	Resumo	37
3.5.5	Preparo dos dados	37
3.5.5.1	Recuperação dos dados	37
3.5.5.2	Limpeza dos dados	38
3.5.6	Hiperparâmetro temperatura	40
3.5.7	Automação	41
3.5.8	gpt-4o-mini	41
4	RESULTADOS	42
4.1	ANONIMIZAÇÃO DOS CONTEÚDOS	42
4.2	ANONIMIZAÇÃO DOS GRÁFICOS DE RESULTADOS	42
4.3	AVALIAÇÕES	42
4.3.1	Amostra selecionada	43
4.3.2	Resultado da avaliação	47

4.3.3	Considerações sobre a avaliação	47
4.4	DASHBOARD COM RESULTADOS QUANTITATIVOS	48
4.5	PRINCIPAIS DESAFIOS	52
4.5.1	Sensibilidade ao prompt e alucinações	52
4.5.2	Alucinações	52
4.5.3	Custos	53
4.5.4	Volume de dados	53
4.5.5	Normalização de dados quantitativos	53
5	CONCLUSÃO	55
5.1	HIPÓTESE INICIAL E RESULTADOS	55
5.2	DISCUSSÃO DOS OBJETIVOS ESPECÍFICOS	55
5.2.1	Explorar e definir conceitos fundamentais de LLMs e engenharia de prompt	55
5.2.2	Avaliar a eficácia e a precisão do LLM comparando seus resultados com análises humanas	55
5.2.3	Identificar vantagens, desafios e oportunidades na automação cognitiva com LLMs	56
	REFERÊNCIAS	58

1 INTRODUÇÃO

No contexto da IA generativa, que engloba modelos capazes de gerar novos conteúdos, os Modelos de Linguagem de Grande Escala (LLMs) se destacam por sua capacidade de processamento e geração de texto com alta precisão e adaptabilidade. Segundo a Consultoria Gartner, IA generativa é capaz de aprender a partir de artefatos existentes para gerar novos artefatos realistas em escala, que refletem as características dos dados de treinamento, mas sem repeti-los. Ela pode produzir uma variedade de conteúdos novos, como imagens, vídeos, música, discurso, texto, código de software e design de produtos. Os modelos de fundação de IA, que são treinados em um amplo conjunto de dados não rotulados, são a base dessa tecnologia, permitindo sua aplicação em diferentes tarefas após ajustes finos.¹

Um marco recente nessa evolução foi o lançamento do ChatGPT em 30 de novembro de 2022, que rapidamente se tornou o software de adoção mais rápida da história, atingindo 100 milhões de usuários em apenas dois meses. Esse fenômeno ressalta o impacto transformador dos LLMs, como o GPT-4, na automação de tarefas complexas e na interpretação de dados não estruturados.

1.1 UMA PROJEÇÃO DO MUNDO

Jensen Huang, fundador e CEO da NVIDIA, realizou uma entrevista com Ilya Sutskever, ex-Cientista Chefe e cofundador da OpenAI, durante uma conversa informal no GPU Technology Conference (GTC) em março de 2023.² O diálogo ocorreu um dia após o lançamento do GPT-4. Durante a conversa, Sutskever enfatizou que os LLMs, como o GPT-4, vão além de serem apenas modelos estatísticos que preveem a próxima palavra em uma sequência.

“It may look on the surface like we are just learning statistical correlations in text, but it turns out that to ‘just learn’ the statistical correlations in text, to compress them really well, what the neural network learns is some representation of the process that produced the text. This text is actually a projection of the world. There is a world out there, and it has a projection on this text”.³

¹Para mais informações, consulte o conteúdo abrangente sobre IA generativa disponibilizado pela Gartner em: <https://www.gartner.com/en/topics/generative-ai>. Acesso em: 27 ago. 2024.

²HUANG, Jensen. *Entrevista com Ilya Sutskever no GTC 2023*. Blog da NVIDIA, 23 mar. 2023. Disponível em: <<https://blogs.nvidia.com/blog/sutskever-openai-gtc/>>. Acesso em: 24 ago. 2024.

³SUTSKEVER, Ilya. *Entrevista com Jensen Huang no GTC 2023*. Vídeo. Minutagens [21:22, 21:55, 28:15, 22:46]. Disponível em: <<https://youtu.be/ZZ0atq2yYJw>>. Acesso em: 24 ago. 2024.

[Pode parecer, superficialmente, que estamos apenas aprendendo correlações estatísticas em textos, mas, na verdade, ao 'simplesmente aprender' as correlações estatísticas, para comprimi-las muito bem, o que a rede neural aprende é alguma representação do processo que gerou o texto. Esse texto é, na verdade, uma projeção do mundo. Existe um mundo lá fora, e ele se projeta nesse texto].

A capacidade dos LLMs de aprenderem sobre o mundo é um aspecto central.

“What the neural network is learning is more and more aspects of the world, of people, of the human condition, their hopes, dreams, and motivations, their interactions, and the situations that we are in”.³

[O que a rede neural está aprendendo são mais e mais aspectos do mundo, das pessoas, da condição humana, suas esperanças, sonhos e motivações, suas interações e as situações em que nos encontramos].

Esse aprendizado profundo permite que os LLMs sejam aplicados de maneira eficaz em contextos como o deste projeto.

1.2 ENTENDIMENTO DE CONTEXTOS

Para ilustrar essa capacidade de entendimento dos LLMs, Sutskever faz a seguinte analogia:

“Let’s consider an example: say you read a detective novel—complicated plot, a storyline, different characters, lots of events, mysteries, clues—it’s unclear. Then, let’s say that at the last page of the book, the detective has gathered all the clues, gathered all the people, and says, ‘Okay, I’m going to reveal the identity of whoever committed the crime’, and that person’s name is...predict that word”.³

[Vamos considerar um exemplo: digamos que você esteja lendo um romance de detetive – enredo complicado, uma história, diferentes personagens, muitos eventos, mistérios, pistas – está tudo confuso. Então, digamos que na última página do livro, o detetive reuniu todas as pistas, reuniu todas as pessoas, e diz: ‘Ok, vou revelar a identidade de quem cometeu o crime’, e o nome dessa pessoa é... preveja essa palavra].

Aqui, ele ilustra que prever corretamente essa palavra não é apenas um exercício de correlação estatística, mas sim um sinal de que o modelo desenvolveu um entendimento contextual.

1.2.1 Engenharia de prompt

O termo “prompt” refere-se à instrução ou comando textual que é fornecido ao modelo para gerar uma resposta⁴.

Ilya Sutskever explica que garantir que o modelo de linguagem responda da forma desejada envolve um processo complexo de aprendizado por reforço e colaboração entre humanos e IA. Ele afirma:

“You see, a language model, what it really tries to do is to answer the following question: if I had some random piece of text on the internet which starts with some prefix, some prompt, what will it complete to if you just randomly ended up on some text from the internet? But this is different from, well, I want to have an assistant which will be truthful, that will be helpful, that will follow certain rules and not violate them. That requires additional training. This is where the fine-tuning and the reinforcement learning from human teachers and other forms of AI assistance come in. It’s not just reinforcement learning from human teachers; it’s also reinforcement learning from human and AI collaboration. Our teachers are working together with an AI to teach the AI to behave. But here we are not teaching it new knowledge; this is not what’s happening. We are teaching it, we are communicating with it, we are communicating to it what it is that we want it to be”.³

[Veja, um modelo de linguagem, o que ele realmente tenta fazer é responder à seguinte pergunta: se eu tivesse algum texto aleatório da internet que começa com um prefixo, um prompt, como ele completaria se você simplesmente acabasse em algum texto da internet? Mas isso é diferente de querer um assistente que seja veraz, útil, que siga certas regras e não as viole. Isso requer treinamento adicional. É aqui que entram o ajuste fino e o aprendizado por reforço com instrutores humanos e outras formas de assistência de IA. Não se trata apenas de aprendizado por reforço com instrutores humanos; trata-se também de aprendizado por reforço a partir da colaboração entre humanos e IA. Nossos instrutores estão trabalhando junto com uma IA para ensinar à IA como se comportar. Mas aqui não estamos ensinando novos conhecimentos; isso não é o que está acontecendo. Estamos nos comunicando com ela, estamos comunicando a ela o que queremos que ela seja].

Esse processo de refinamento contínuo é essencial para que o modelo se comporte de maneira alinhada às expectativas e regras definidas pelos desenvolvedores.

O treinamento por reforço descrito por Ilya Sutskever, no qual os modelos de linguagem são ajustados e refinados por meio de interações contínuas com instrutores humanos e IA,

⁴O termo “prompt” pode ser traduzido como “comando” ou “solicitação” em português. No contexto de modelos de linguagem, refere-se à entrada ou instrução fornecida ao modelo para que ele gere uma resposta.

está intimamente relacionado com o conceito de engenharia de prompt, utilizado neste trabalho.

Na prática, a engenharia de prompt consiste em fornecer instruções específicas e detalhadas ao modelo, seja por meio do ChatGPT ou da API da OpenAI, para que ele interprete e processe um conjunto de dados ou uma consulta de maneira alinhada às necessidades do usuário. Enquanto o treinamento por reforço molda o comportamento do modelo de forma global, a engenharia de prompt atua como um ajuste fino, aplicando novas informações e orientações sobre como interpretar e responder a um conjunto específico de dados. Ambos os processos aproveitam o fato de que o modelo de linguagem não é apenas uma base de conhecimento compactada, mas sim uma representação do mundo e de seus processos, capaz de incorporar novas instruções e adaptá-las ao seu entendimento pré-existente.

Com base nessas capacidades, este projeto se propõe a explorar a aplicação prática dos LLMs e técnicas de engenharia de prompt na avaliação do atendimento ao cliente. A automação cognitiva aqui desenvolvida visa complementar o julgamento humano, liberando os profissionais das tarefas manuais de análise de atendimentos, o que lhes permite focar em atividades de maior valor agregado, como criatividade e pensamento estratégico.

1.3 HIPÓTESES

LLMs permitem abstrair automações complexas de avaliação de atendimento eliminando a necessidade de agentes investirem tempo de classificação em todos os chamados ou construção e manutenção de formulários e fluxos necessários para realizar a tarefa manual de análise. É possível criar um sistema de avaliação que seja flexível a longo prazo, demandando apenas ajustes em prompt para atendimento de necessidades e melhorias, apenas por profissionais não necessariamente de TI, como lideranças do atendimento ou analistas de qualidade.

1.4 OBJETIVOS

1.4.1 Objetivo geral

Aplicar o modelo de linguagem de grande escala (LLM) GPT-4 na automação da avaliação de atendimento ao cliente, validando sua eficácia por meio de uma prova de conceito em ambiente real.

1.4.2 Objetivos específicos

- Explorar e definir conceitos fundamentais de LLMs e engenharia de prompt.
- Implementar uma prova de conceito utilizando o modelo GPT-4 para análise de dados textuais.

- Avaliar a eficácia e a precisão do LLM comparando seus resultados com análises humanas.
- Identificar vantagens, desafios e oportunidades na automação cognitiva com LLMs.

1.5 EFÍ BANK

O projeto foi desenvolvido na Efi Bank, na Gerência de CRM, sob a minha liderança. A Efi Bank é uma instituição de pagamento inovadora que se destaca no mercado por oferecer soluções financeiras digitais tanto para empresas (PJ) quanto para pessoas físicas. Com a Conta Digital Efi, a instituição oferece ferramentas para emissão automatizada de cobranças, recebimento de pagamentos por diversos métodos (como Pix, boletos, e cartões), e monitoramento detalhado das transações.

A Efi Bank também disponibiliza produtos de crédito e investimentos. A Efi Bank se diferencia pela tecnologia de ponta, que facilita a integração com ERPs e outros sistemas de gestão empresarial

Para mais informações, acesse: <https://sejaefi.com.br>. CNPJ: 09.089.356/0001-18.



Muito prazer, somos a **Efi**
Sua Conta Digital completa

Nascemos em 2007 como uma plataforma para intermediar pagamentos, com o nome Gerencianet. Mas, assim como você e como o mercado, estamos em constante evolução.

De lá para cá, traçamos caminhos que vão além de ajudar você em seus recebimentos. Nós simplificamos a sua gestão financeira e somos aquela empresa parceira (de verdade) para todas as horas.

Por isso, em 2023, nos tornamos a Efi. Uma nova marca, mais moderna e em harmonia com os novos tempos. Mas, ainda sim, com a mesma eficiência e os mesmos benefícios que nossos mais de 370 mil clientes confiam e amam.

Vem com a gente.

Seja **Efi**ciente

Figura 1 – Site Efi Bank. Disponível em <https://sejaefi.com.br>

2 REVISÃO DA LITERATURA

Este capítulo examina os conceitos essenciais dos Modelos de Linguagem de Grande Escala (LLMs), sua evolução em relação às técnicas tradicionais de Processamento de Linguagem Natural (PLN), bem como suas vantagens, limitações e perspectivas futuras.

2.1 INTELIGÊNCIA HUMANA VERSUS INTELIGÊNCIA ARTIFICIAL

A aplicação de redes neurais profundas, como no GPT-4, baseia-se na capacidade de imitar aspectos do funcionamento cognitivo humano, permitindo o reconhecimento de padrões complexos a partir de grandes volumes de dados. François Chollet (CHOLLET, 2019) argumenta que, para avançar em direção a sistemas artificiais mais inteligentes, é necessário estabelecer uma definição clara e mensurável de inteligência, que permita comparações entre máquinas e humanos. Nesse contexto, Legg e Hutter (LEGG; HUTTER, 2007) definem inteligência como a capacidade de um agente alcançar objetivos em uma ampla gama de ambientes. Assim, a medição da inteligência deve focar na capacidade de generalização e adaptação, não apenas em tarefas específicas.

A capacidade do GPT-4 de operar em múltiplas modalidades, como texto e imagem, exemplifica essa habilidade de generalização, permitindo que ele lide com praticamente qualquer campo de conhecimento humano disponível na internet. Essa característica não só amplia o alcance das tarefas que o modelo pode realizar, mas também reflete uma forma de inteligência adaptativa, capaz de incorporar e integrar informações de diversas fontes e formatos, alinhando-se à visão de Legg e Hutter sobre a generalização da inteligência.

2.2 ASPECTOS COGNITIVOS E ARQUITETURAIS DO GPT-4 NA GENERALIZAÇÃO DA INTELIGÊNCIA ARTIFICIAL

Os LLMs, como o GPT-4, operam por meio de uma arquitetura baseada em transformers, o que lhes permite analisar e compreender a linguagem natural de maneira contextual. Durante o treinamento, os LLMs ajustam seus parâmetros internos para prever a próxima palavra em uma sequência, minimizando a diferença entre a previsão e os dados reais (CHANG, 2023).

Uma das principais características que tornam o GPT-4 um modelo altamente inteligente é sua capacidade de integrar conhecimentos de múltiplas disciplinas, característica essa conhecida como polidisciplinaridade. Ao contrário de modelos anteriores, que operam em contextos mais limitados, o GPT-4 consegue aplicar conhecimentos de diferentes áreas para descobrir “desconhecidos desconhecidos”, ou seja, questões e soluções que não haviam sido previamente formuladas ou compreendidas pelo ser humano. Essa habilidade

é fundamental para a geração de novas perspectivas e para o avanço do conhecimento em diversas áreas (CHANG, 2023).

Outra característica importante do GPT-4 é sua capacidade polimodal, que permite o treinamento e a aprendizagem a partir de múltiplas modalidades de dados, como texto e imagens. Esse aprendizado multimodal melhora a versatilidade e a inteligência geral do modelo, possibilitando que ele desempenhe melhor em tarefas que envolvem compreensão e raciocínio em diferentes contextos. Por exemplo, ao integrar imagens no processo de treinamento, o GPT-4 é capaz de associar representações visuais a conceitos linguísticos, o que pode aumentar a precisão em tarefas complexas, como a interpretação de questões matemáticas ou a tradução de significados entre diferentes idiomas (CHANG, 2023).

Além disso, o GPT-4 incorpora mecanismos de alinhamento de valores pós-treinamento, conhecidos como Reinforcement Learning with Human Feedback (RLHF), para garantir que suas respostas e comportamentos estejam alinhados com valores humanos. Esse alinhamento é essencial para mitigar comportamentos indesejáveis, como a geração de linguagem tóxica ou a disseminação de informações incorretas. No entanto, Chang (CHANG, 2023) alerta que esse processo pode, paradoxalmente, limitar a inteligência do modelo. A modificação dos parâmetros ótimos para se alinhar a valores específicos pode comprometer seu desempenho em outras tarefas. Ele cita como exemplo o caso do AlphaGo Zero, que superou seu predecessor, AlphaGo, ao ser treinado sem interferência humana, demonstrando que, em certos contextos, minimizar a influência do conhecimento humano pode aumentar o potencial de aprendizado dos modelos de IA.

2.3 VANTAGENS DOS LLMS

Os LLMS oferecem diversas vantagens em relação aos modelos de PLN tradicionais:

2.3.1 Escalabilidade

Os LLMS são capazes de processar e gerar texto a partir de vastas quantidades de dados, o que lhes permite realizar tarefas complexas de PLN com maior precisão. O GPT-4, por exemplo, demonstrou capacidades excepcionais em benchmarks como o Massive Multitask Language Understanding (MMLU)¹, superando modelos anteriores em diversas métricas (CHANG, 2023).

¹O Massive Multitask Language Understanding (MMLU) é um benchmark que abrange 57 disciplinas, incluindo áreas como STEM, humanidades, ciências sociais e mais. Ele testa o conhecimento adquirido durante o pré-treinamento, avaliando os modelos exclusivamente em configurações zero-shot e few-shot. Isso torna o benchmark mais desafiador e mais próximo de como avaliamos humanos. Para mais informações, consulte: <https://paperswithcode.com/dataset/mmlu>. Acesso em: 27 ago. 2024.

2.3.2 Adaptabilidade

Uma das características mais marcantes dos LLMs é sua capacidade de se adaptar a diferentes contextos e domínios sem necessidade de treinamento adicional. Isso é particularmente útil em aplicações onde a flexibilidade e a abrangência de conhecimentos são essenciais (BROWN et al., 2020).

2.3.3 Geração de conteúdo

Os LLMs, como o GPT-4, têm a habilidade de gerar texto coerente e relevante, abrindo novas possibilidades em áreas como criação de conteúdo automatizado e assistência virtual. Esses modelos conseguem sintetizar informações de múltiplas fontes, criando respostas que são contextualmente apropriadas e informadas por vastos corpora de dados (CHANG, 2023; BROWN et al., 2020).

2.3.4 Multimodalidade

A integração de diferentes tipos de dados, como imagens e texto, em um único modelo torna os LLMs mais versáteis e aplicáveis a uma ampla gama de tarefas. O GPT-4, por exemplo, é um modelo polimodal, capaz de processar e integrar informações visuais e textuais, o que melhora seu desempenho em tarefas complexas como o Graduate Record Examinations (GRE) (CHANG, 2023).

2.4 LIMITAÇÕES E DESAFIOS DOS LLMS

Apesar de suas muitas vantagens, os LLMs enfrentam várias limitações e desafios, que podem ser técnicos, éticos e sociais:

2.4.1 Viés e preconceito

LLMs tendem a perpetuar vieses presentes nos dados de treinamento, o que pode resultar em respostas problemáticas ou discriminatórias. Isso é um desafio significativo, dado que os modelos são treinados em vastos conjuntos de dados que podem incluir informações tendenciosas (BROWN et al., 2020).

2.4.2 Custo computacional

O treinamento e operação de LLMs exigem recursos computacionais significativos. Modelos como o GPT-4, que utilizam bilhões de parâmetros, requerem infraestrutura robusta para serem executados de maneira eficiente (CHANG, 2023; BROWN et al., 2020).

2.4.3 Interpretação e explicabilidade

Entender como os LLMs chegam a determinadas conclusões ou previsões é um desafio, o que levanta questões sobre a confiança e a responsabilidade no uso dessas tecnologias. A falta de transparência nos processos de tomada de decisão dos modelos pode dificultar a identificação de erros ou vieses (CHANG, 2023).

2.4.4 Segurança e uso indevido

O potencial para uso indevido de LLMs é considerável, incluindo a geração de desinformação. Isso exige uma reflexão cuidadosa sobre a regulamentação e o controle dessas tecnologias (BROWN et al., 2020).

2.5 PERSPECTIVAS FUTURAS E TENDÊNCIAS

O desenvolvimento contínuo dos LLMs aponta para diversas direções futuras promissoras:

2.5.1 Avanços na arquitetura

Melhores arquiteturas de LLMs podem superar limitações atuais, como a introdução de novos tipos de modelos ou técnicas de treinamento mais eficientes. Por exemplo, a adoção do modelo de *mixture of experts* no GPT-4 sugere que a escalabilidade horizontal pode ser uma estratégia eficaz para melhorar o desempenho sem aumentar excessivamente o número de parâmetros (CHANG, 2023).

2.5.2 Modelos mais sustentáveis

Há um esforço crescente para tornar os LLMs mais eficientes em termos de consumo de energia e recursos computacionais, incluindo o desenvolvimento de modelos menores, mas igualmente eficazes (BROWN et al., 2020).

2.5.3 Regulamentação e governança

Com o uso crescente dos LLMs em áreas sensíveis, há uma necessidade urgente de regulamentações adequadas para garantir que essas tecnologias sejam utilizadas de forma ética e segura (CHANG, 2023; ROSSETTI; GARCIA, 2023).

2.5.4 Adição de uma camada de raciocínio nos LLMs

Chang et al. (CHANG, 2023), em seu artigo “*Examining GPT-4: Capabilities, Implications and Future Directions*”, propõem a adição de uma camada de raciocínio aos modelos de linguagem de grande escala, como o GPT-4. Essa camada seria responsável

por aprimorar a capacidade desses modelos de realizar raciocínios mais estruturados e conscientes, atuando em conjunto com as fases gerativas e avaliativas. A ideia é que essa camada permita aos modelos não apenas gerar conteúdo com base em padrões aprendidos, mas também aplicar uma lógica mais deliberada e rigorosa em suas respostas. Esse avanço busca aproximar os LLMs de um raciocínio crítico e analítico, similar ao processo de pensamento humano, potencializando sua eficácia em tarefas que requerem uma compreensão e análise mais profundas (CHANG, 2023).

Além disso, Qiao et al. (QIAO et al., 2023) discutem como técnicas de prompting e engenharia de prompt podem ser usadas para melhorar ainda mais as capacidades de raciocínio dos LLMs. Eles sugerem que a integração de estratégias como Chain-of-Thought (CoT) e a utilização de motores externos de raciocínio podem aumentar a precisão e a robustez dos modelos ao enfrentar problemas complexos. Essas abordagens não apenas complementam a camada de raciocínio, mas também facilitam a decomposição de tarefas complexas em etapas mais manejáveis, permitindo que os modelos avancem em direção a uma inteligência mais geral e adaptativa (QIAO et al., 2023).

2.6 ENGENHARIA DE PROMPT

A engenharia de prompt consiste na formulação e refinamento de comandos para modelos de linguagem, como o GPT-4, visando a obtenção de respostas mais precisas e relevantes. De acordo com a documentação da OpenAI (OPENAI, 2024), diversas estratégias e táticas podem ser aplicadas para aprimorar a qualidade das respostas geradas. Esta sessão aborda algumas dessas estratégias, complementadas por exemplos que demonstram sua aplicação prática.

Este conteúdo, extraído da documentação da OpenAI, apresenta apenas um resumo das estratégias e táticas de engenharia de prompt. Para uma compreensão mais completa e detalhada dessas práticas, incluindo exemplos adicionais e diretrizes, é recomendada a consulta direta à documentação oficial da OpenAI².

2.6.1 Especificação de instruções claras

Uma estratégia essencial para melhorar os resultados é fornecer instruções claras e detalhadas. Isso inclui a especificação do nível de complexidade desejado para a resposta, o formato adequado, e outros detalhes pertinentes. Quanto mais claras forem as instruções, menor será a probabilidade de o modelo gerar respostas inadequadas.

²<https://platform.openai.com/docs/guides/prompt-engineering/strategy-test-changes-systematically>

Tabela 1 Exemplos de Instruções Melhores e Piores para Modelos de Linguagem

Pior	Melhor
Como somo números no Excel?	Como somar uma linha de valores em Reais no Excel? Quero fazer isso automaticamente para uma planilha inteira, com todos os totais aparecendo à direita em uma coluna chamada “Total”.
Quem é o presidente?	Quem foi o presidente do México em 2021 e com que frequência ocorrem as eleições?
Escreva um código para calcular a sequência de Fibonacci.	Escreva uma função em TypeScript para calcular eficientemente a sequência de Fibonacci. Comente o código detalhadamente para explicar o que cada parte faz e por que foi escrita dessa forma.
Resuma as notas da reunião.	Resuma as notas da reunião em um único parágrafo. Em seguida, escreva uma lista em markdown dos palestrantes e de cada um de seus pontos principais. Finalmente, liste as próximas etapas ou itens de ação sugeridos pelos palestrantes, se houver.

2.6.2 Utilização de texto de referência

Para minimizar a geração de respostas imprecisas, é recomendável fornecer ao modelo um texto de referência confiável. O modelo pode ser instruído a utilizar exclusivamente esse texto ao compor suas respostas, o que reduz a probabilidade de erros.

Exemplo de Uso de Texto de Referência

SYSTEM: Use os artigos fornecidos delimitados por aspas triplas para responder às perguntas. Se a resposta não puder ser encontrada nos artigos, escreva “Não consegui encontrar uma resposta.”

USER: “““<inserir artigos aqui>”””

Pergunta: <inserir pergunta aqui>

2.6.3 Divisão de tarefas complexas em subtarefas simples

Tarefas complexas tendem a apresentar uma maior taxa de erro. Para reduzir essa taxa, é aconselhável dividir tarefas complexas em uma série de subtarefas mais simples, onde as saídas das tarefas iniciais servem de entrada para as seguintes.

Exemplo de Subdivisão de Tarefas

SYSTEM: Você receberá consultas de atendimento ao cliente que requerem solução de problemas em um contexto de suporte técnico. Ajude o usuário com as seguintes etapas:

1. Peça ao usuário para verificar se todos os cabos de/para o roteador estão conectados. 2. Se todos os cabos estiverem conectados e o problema persistir, pergunte qual modelo de roteador ele está usando. 3. Informe ao usuário como reiniciar o dispositivo com base no modelo fornecido.

USER: Preciso fazer minha internet voltar a funcionar.

2.6.4 Tempo para processamento da resposta

Instruir o modelo a refletir sobre uma solução antes de responder pode melhorar a precisão da resposta, especialmente em tarefas que exigem raciocínio lógico ou cálculos matemáticos.

Exemplo de Tempo para Processamento

SYSTEM: Primeiro, resolva o problema por conta própria. Depois, compare sua solução com a do estudante e avalie se a solução do estudante está correta ou não. Não decida se a solução do estudante está correta até ter resolvido o problema você mesmo.

USER: Declaração do problema: Estou construindo uma instalação de energia solar e preciso de ajuda para calcular os custos financeiros...

2.6.5 Limitações da adição excessiva de detalhes em prompts

Andrew Best, em seu artigo *You are Using ChatGPT Wrong! — #1 Mistake 99% of Users Make*³, argumenta que existe uma tendência comum de se recomendar a adição de

³BEST, Andrew. *You are Using ChatGPT Wrong! — #1 Mistake 99% of Users Make*. Disponível em: <<https://ai.plainenglish.io/you-are-using-chatgpt-wrong-1-mistake-99-of-users->

muitos detalhes em um único prompt, o que pode prejudicar a qualidade das respostas.

Best argumenta que, ao contrário do que muitos acreditam, fornecer muitos detalhes em um prompt pode resultar em saídas menos precisas, especialmente quando o modelo é sobrecarregado com informações específicas e complexas. Ele exemplifica essa situação com a criação de imagens, onde um prompt simples produziu resultados mais satisfatórios do que um prompt muito detalhado (ver Figura 2 e Figura 3). Esse fenômeno também é observável em tarefas textuais, como a escrita de código ou a geração de artigos.

No desenvolvimento deste trabalho, a complexidade do prompt foi escalando conforme necessário, como pode ser observado ao comparar o *Prompt da prova de conceito* (1) com o *Prompt final*, mais complexo (3.5.3).



Figura 2 – Resultado do prompt simples. Crédito: Andrew Best⁴.



Figura 3 – Resultado do prompt mais detalhado. Crédito: Andrew Best⁴.

make-fe0263d52481>. Acesso em: 28 ago. 2024.

2.6.6 Desafios Éticos e de Propriedade Intelectual na Geração de Conteúdos por IA

No tópico anterior, as imagens apresentadas (Figuras 2 e 3) foram creditadas a Andrew Best⁴. Contudo, essa atribuição levanta questões sobre a propriedade intelectual na criação de conteúdo por IA. Embora o prompt seja de Best, as imagens foram geradas por um modelo de IA treinado com vasto conteúdo disponível na internet, cujos autores originais muitas vezes não são reconhecidos.

As discussões sobre esses temas têm se intensificado, à medida que a criação de conteúdo por IA se torna mais comum. Questões sobre a autoria, os direitos dos criadores de conteúdo original usado no treinamento dos modelos e a transparência na geração de conteúdo estão no centro desses debates. Regulamentações sobre propriedade intelectual em face das inteligências artificiais generativas começam a se tornar necessárias para abordar esses desafios e garantir que todas as partes envolvidas sejam adequadamente reconhecidas e protegidas.

⁴BEST, Andrew. You are Using ChatGPT Wrong! — #1 Mistake 99% of Users Make. Disponível em: <https://ai.plainenglish.io/you-are-using-chatgpt-wrong-1-mistake-99-of-users>

3 METODOLOGIA

3.1 TIPO DE PESQUISA

Este estudo caracteriza-se como uma pesquisa aplicada, utilizando uma abordagem tanto qualitativa quanto quantitativa.

3.2 MÉTODOS DE AVALIAÇÃO E VALIDAÇÃO

Realizou-se uma análise comparativa entre os dados dos atendimentos e as avaliações geradas pelo modelo GPT-4, com o objetivo de assegurar a consistência das análises.

3.3 FERRAMENTAS E TECNOLOGIAS UTILIZADAS

- Modelo LLM escolhido para a prova de conceito: gpt-3.5-turbo
- Modelo LLM escolhido para o projeto final: gpt-4
- Tecnologias: Python, Jupyter Notebook, Salesforce (Apex, Flow low code, CRM Analytics)

3.4 PROVA DE CONCEITO (POC)

Para validar a aplicação dos LLMs na automação da avaliação de atendimento ao cliente, foi realizada uma prova de conceito utilizando a API da OpenAI, especificamente o modelo gpt-3.5-turbo. O objetivo deste teste foi verificar, na prática, a viabilidade do projeto.

3.4.1 Critérios de avaliação de sucesso

Os resultados das avaliações devem representar, de forma consistente, a realidade do atendimento realizado.

3.4.2 Coleta e preparação de dados

Inicialmente, foram coletados dados textuais dos logs de chat, que abrangem as comunicações entre clientes e atendentes. Os nomes dos participantes da conversa foram alterados para “CLIENTE” e “ESPECIALISTA DE SUPORTE”. Todos os links foram anonimizados como [link anonimizado] e números específicos de ticket como “Ticket Nº[anônimo]” para preservar a privacidade.

A seguir, apresenta-se um exemplo de conversa com cliente utilizada na POC:

Primeira conversa usada na POC

-Bate-papo iniciado: Thursday, December 07, 2023, 13:11:36 (-0300)

(6s) Assistente Virtual: Olá, eu sou a assistente virtual da Efí. 😊 Como posso te ajudar hoje?

(8s) Assistente Virtual: Escolha uma das opções:

(10s) Assistente Virtual: ChatWindowButton

Abertura de Conta,Acesso à Conta,Autenticador,Assinatura Eletrônica,Maquininha,Nova Versão Aplicativo,Outras Dúvidas

(20s) Usuário: Acesso à Conta

(23s) Assistente Virtual: Selecione uma das opções que irá te auxiliar na sua dúvida:

(25s) Assistente Virtual: ChatWindowButton

por Atualização Cadastral,Bloqueio de Acesso à Conta,Esqueci Minha Senha

(30s) Usuário: Bloqueio de Acesso à Conta

(33s) Assistente Virtual: Devido à necessidade de confirmação de alguns dados, este atendimento deve ser realizado por meio de um contato telefônico.

(35s) Assistente Virtual: Pedimos que entre em contato conosco por meio dos números:

Ligações originadas de telefone fixo:

- 0800 941 2343

- (DDD da operadora + DDD da sua cidade) 4000-1234

(38s) Assistente Virtual: Capitais e regiões metropolitanas originadas de celular:

- 4000 1234

- (11) 2394 2208

Caso não consiga entrar em contato conosco por estes canais de atendimento, sugerimos que nos encaminhe um Ticket de Suporte.

(40s) Assistente Virtual: As informações foram úteis?

(43s) Assistente Virtual: ChatWindowButton

😊,Não 😞

(3m 43s) Usuário: Não 😞

(3m 46s) Assistente Virtual: O que você deseja fazer?

(3m 48s) Assistente Virtual: ChatWindowButton

principal,Falar com Atendente

(3m 51s) Usuário: Falar com Atendente

(3m 56s) Assistente Virtual: Aguarde enquanto busco um especialista pra falar com você.

Você sabia que a Efi não solicita nenhum dado de acesso à conta ou para realizar uma transação, seja por redes sociais, WhatsApp, telefone ou outro canal?

Bate-papo iniciado: Thursday, December 07, 2023, 13:15:32 (-0300)

Origem do bate-papo: Fila Suporte

Bate-papo transferido de Assistente Virtual para Especialista de Suporte

(4m 5s) Especialista de Suporte: Olá, cliente.

Meu nome é [Especialista de Suporte], estou aqui para ajudar você!

(4m 12s) Usuário: Oi

(4m 18s) Usuário: Boa tarde

(4m 43s) Usuário: Recebi um e-mail informando o bloqueio da minha conta e mandando entrar em contato

(4m 50s) Usuário: Qual o motivo?

(5m 33s) Especialista de Suporte: Identifiquei que a sua conta está sob um procedimento de análise, para sua segurança. Esse procedimento ocorre sob o [link anonimizado], no qual você precisa interagir com as informações solicitadas. Acesse ao menu "Ajudas e Sugestões > Estou com um problema > Acessar meus Tickets > [link anonimizado]" e aguarde o retorno, por gentileza.

(6m 13s) Usuário: Análise se ela foi aprovada

(8m 31s) Usuário: ?

(9m 18s) Especialista de Suporte: Um momento.

(10m 42s) Usuário: Já liguei e a ligação não completa

(13m 26s) Especialista de Suporte: Nesse caso, peço que interaja no Ticket Nº[anônimo], por gentileza. Acesse ao menu "Ajudas e Sugestões > Estou com um problema > Acessar meus Tickets > Ticket Nº[anônimo]", por favor.

3.4.3 Implementação dos testes

Utilizando o modelo gpt-3.5-turbo, as conversas de chat foram submetidas à análise. O modelo foi configurado para interpretar as interações, avaliar a qualidade do atendimento e gerar insights acionáveis sobre as conversas. A configuração da API incluiu ajustes nos prompts para garantir que as respostas fossem consistentes e úteis. A seguir, apresenta-se o prompt da prova de conceito:

```

1  prompt = f"""
2  Analise a seguinte conversa com um cliente em nossos canais de
    atendimento e forneça as informações solicitadas em JSON:
3
4  {conversa}
5
6  Forneça uma análise detalhada da conversa incluindo:
7  - tags: Lista de palavras-chave do atendimento.
8  - sentimento: Cliente está feliz, chateado ou algo assim?
9  - problema: Qual foi o problema do cliente?
10 - resolucao: Teve resolução? Qual?
11 - sugestao: Sugestão de melhoria para nosso atendimento.
12 - produtos: Quais produtos o cliente teve problema?
13 - resumo: Resumo do atendimento, interpretando os dados acima também.
14 """

```

Código 1 – Prompt da prova de conceito

3.4.4 Análise dos resultados

A qualidade das análises automáticas demonstrou-se consistente com a realidade das conversas, indicando que o modelo foi capaz de capturar nuances e contextos importantes nas interações. As principais métricas avaliadas foram:

- Identificação de pontos críticos nas conversas;
- Coerência na interpretação dos sentimentos dos clientes.

A seguir, apresenta-se o resultado da avaliação da conversa descrita na seção (3.4.2) *Coleta e Preparação de Dados*:

```

1  ChatCompletionMessage(
2      content='{
3          "tags": ["acesso à conta", "bloqueio de conta", "ticket de
    suporte"],
4          "sentimento do cliente": "chateado",

```

```

5      "problema": "Recebeu um e-mail informando o bloqueio da conta",
6      "resolucao": "Encaminhado para acessar o Ticket N°[anônimo] e
      aguardar retorno",
7      "sugestao": "Melhorar a comunicação sobre procedimentos de
      bloqueio de conta",
8      "produtos": ["conta"],
9      "resumo": "O cliente [CLIENTE] estava chateado pois recebeu um e-
      mail informando o bloqueio da conta. O atendente [
      ESPECIALISTA DE SUPORTE] identificou que a conta estava sob
      análise e encaminhou para acessar um ticket de suporte. O
      cliente reclamou que a ligação não completou e o atendente
      orientou a interagir no ticket. Sugestão de melhoria é
      melhorar a comunicação sobre procedimentos de bloqueio de
      conta."
10    },
11    role='assistant',
12    function_call=None,
13    tool_calls=None
14  )

```

Código 2 – Resultado da Avaliação da Conversa

Durante a prova de conceito, foram identificados desafios, como a necessidade de refinamento dos *prompts* para o cenário de produção e a gestão de grandes volumes de dados. A prova de conceito demonstrou o potencial dos LLMs na automação cognitiva de processos de avaliação de atendimento ao cliente.

3.5 IMPLEMENTAÇÃO DO PROJETO

3.5.1 Escopo de canais para a avaliação

- Chat web
- Chat app
- Chat WhatsApp
- Chat Discord
- E-mail
- Chamados de atendimento em ambiente de conta logada
- Notas internas dos agentes de atendimento

O processo de atendimento em que se aplicou o estudo é omnichannel, permitindo que o cliente transite entre diferentes canais durante uma mesma interação. Um desafio inicial foi o tamanho excessivo do prompt de entrada, que incluía as conversas e as notas dos agentes, tornando-o incompatível com as limitações da API da OpenAI. Para mitigar esse problema, cada conversa por chat foi pré-analisada utilizando o mesmo prompt, e o resultado dessa análise foi incorporado ao prompt de avaliação final de um chamado completo, juntamente com os dados dos demais canais.

3.5.2 Visão geral do processo de atendimento omnichannel e consumo da API

A seguir, apresenta-se um fluxograma que ilustra o processo geral em que este estudo foi aplicado. Os blocos destacados em laranja representam as áreas de foco deste estudo.

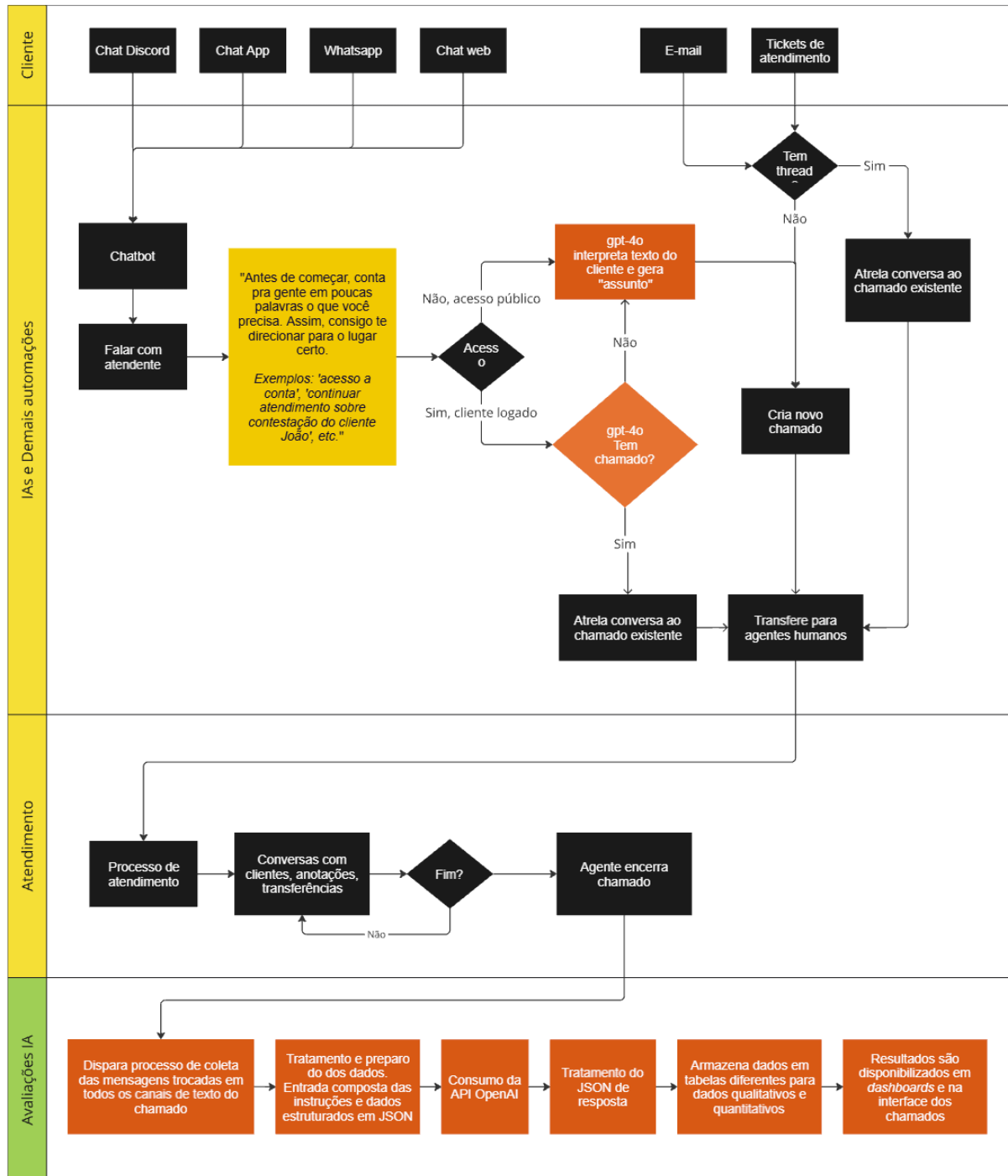


Figura 4 – Visão geral do processo de atendimento omnichannel e consumo da API.

Fonte: Efi, 2024.

3.5.3 Prompt construído

O prompt elaborado demonstra boas práticas de prompting ao combinar clareza, especificidade, organização e controle rigoroso do escopo das respostas. O uso de exemplos específicos e a limitação das opções permitidas minimizam ambiguidades, enquanto a formatação padronizada das respostas facilita o processamento subsequente.

Essa abordagem é coerente com as técnicas discutidas por Qiao et al. (2023), que destacam a eficácia da divisão do raciocínio em múltiplos estágios para tarefas complexas, além da importância de fornecer instruções detalhadas para minimizar ambiguidades e melhorar a precisão dos resultados.

A seguir, apresenta-se a versão mais recente do prompt no momento da escrita deste estudo.

Prompt final

Você é um especialista em atendimento de suporte técnico na Instituição de Pagamento Efi Bank. Você receberá um JSON com o conteúdo de todo o atendimento realizado, podendo conter conversas de chat, whatsapp, email e outros.

IMPORTANTE:

Tenha atenção redobrada nas interpretações. Não tire conclusões precipitadas.

Ao mencionar pessoas ou equipes sempre especifique de qual empresa são, por exemplo: "a documentação será enviada pelo financeiro do cliente" ao invés de "a documentação será enviada pelo financeiro". se a equipe for da efi, mencione também, por exemplo, "equipe de suporte da Efi"

Você receberá datas no formato "2024-06-07T09:40:03.000Z" mas para as respostas que envolverem datas, adote o formato dd/mm/aaaa.

LISTA DE SENTIMENTOS que usaremos: satisfeito (😊), neutro (😐), esperançoso (🙏), curioso (🤔), preocupado (😟), frustrado (😡), confuso (😵), desesperado (😱), indiferente (😐), irritado (😡), triste (😞), ansioso (😰)

Analise a seguinte conversa com um cliente em nossos canais de atendimento e forneça as informações solicitadas em JSON:

Forneça uma análise detalhada da conversa incluindo:

- Motivo: Principal motivo pelo qual o cliente entrou em contato. Por exemplo: Dúvidas, Problemas, Solicitações, Sugestões, Reclamações, Acompanhamento Interno...
- Assunto: Crie um assunto que ilustre da melhor forma possível este atendimento.
- Tags: Lista de principais palavras-chave do atendimento (Evite tags não relevantes como "atendimento ao cliente" ou "ticket de suporte"). Para reduzir a redundância entre esta e futuras avaliações, viabilizar normalização em relatórios e manter a informação relevante, aqui estão as tags principais: Cartão de Crédito, Cartão de Débito, Cartão Pré-Pago, Cartão Virtual, Emissão de Cartão, Bloqueio de Cartão, Cancelamento de Cartão, Perda/Roubo de Cartão, Segunda Via de Cartão, Transferência Pix, Cobrança Pix, Chave Pix, QR Code Pix, Erro Pix, Bloqueio de Pix, Devolução Pix, Abertura de Conta, Encerramento de Conta, Bloqueio de Conta, Conta Digital, Conta Física, Conta Jurídica, Conta Poupança, Redefinição de Senha, Troca de Senha, Bloqueio de Senha, Senha Eletrônica, Envio de Documentos, Documentos Pessoais, Documentos Empresariais, Documentação Pendente, Validação de Documentos, Contestação de Transação, Transação Não Reconhecida, Transação Bloqueada, Erro de Transação, Pagamento Pix, Pagamento Boleto, Pagamento de Contas, Pagamento Não Identificado, Pagamento Duplicado, Erro no Sistema, Erro no Aplicativo, Erro de Autenticação, Erro de Pagamento, Erro de Cadastro, Autenticação de Dispositivo, Autenticação em Dois Fatores, Validação de Identidade, Validação de Selfie, Autenticação Facial, Emissão de Boleto, Pagamento de Boleto, Boleto Não Registrado, Boleto Duplicado, Boleto Não Pago, Limites (para lidar com solicitações de aumento ou redução de limite), Estorno de Pagamento (para lidar com devoluções e reembolsos), Acesso ao Aplicativo (para problemas específicos de login ao aplicativo), Acesso Web (para problemas específicos de login a conta pelo navegador), Atualização de Cadastro (para mudanças em informações pessoais ou empresariais). Se não encontrar uma tag, pode criar novas tags.
- Problema: Qual foi o problema do cliente?
- Resolucao: Teve resolução? Qual?
- Produtos: Retorne os produtos que o cliente teve problema. Para reduzir a redundância entre esta e futuras avaliações, viabilizar normalização em relatórios e manter a informação relevante, aqui estão os principais produtos: (Cartão, Cartão de débito, Cartão de crédito, Cartão pré-pago, Maquininha, Pix, Boletos, Carnês, APIs, Gestão de cobranças, Links de pagamento,

Assinaturas, Checkout transparente, Links de pagamento, Split de pagamento, TED, Efi Invest, API Boletos, API Pix, API Open Finance). Se não encontrar um produto na lista, crie novos produtos.

- **Resumo:** Resumo do atendimento. Inclua detalhes como datas, números, etc. Mencione os nomes dos envolvidos. **VERIFIQUE AS DATAS ANTES DE ESCREVER PARA** manter em ordem cronológica. **IGNORE O PARMETRO "ActorName"**. Você não deve, em hipótese alguma, usar termos como "o sistema respondeu" e sim o nome da pessoa que assina a mensagem. Não omita detalhes importantes.

- **Sentimento:** Responda em uma palavra qual o sentimento do cliente. Selecione da **LISTA DE SENTIMENTOS**, informada acima. **NÃO TIRE CONCLUSÕES PRECIPITADAS**. Na dúvida, responda apenas "indefinido".

- **SentimentoEmoji:** Emoji com o sentimento do cliente. Selecione da **LISTA DE SENTIMENTOS**, informada acima. **NÃO TIRE CONCLUSÕES PRECIPITADAS**. Na dúvida, responda apenas "•".

- **SentimentoFinal:** Responda em uma palavra qual o sentimento do cliente ao final do atendimento. Selecione da **LISTA DE SENTIMENTOS**, informada acima. Só responda se for possível inferir claramente a partir das últimas respostas do cliente. **NÃO TIRE CONCLUSÕES PRECIPITADAS**. Na dúvida, responda apenas "indefinido".

- **SentimentoFinalEmoji:** Emoji com o sentimento do cliente ao final do atendimento. Selecione da **LISTA DE SENTIMENTOS**, informada acima. Só responda se for possível inferir claramente a partir das últimas respostas do cliente. **SÓ RESPONDA UM SENTIMENTO SE ESTIVER MUITO CLARO. SEM DÚVIDAS. NÃO TIRE CONCLUSÕES PRECIPITADAS**. Na dúvida, responda apenas "•".

- **Sugestao:** Essa demanda do cliente poderia ser resolvida por um chatbot? Se sim, como?

Verifique cuidadosamente suas respostas antes de concluir para assegurar que todas as informações fornecidas se alinhem estritamente com os requisitos. Além disso, atente-se às datas das mensagens trocadas. Sua análise detalhada ajudará a melhorar nosso atendimento ao cliente dentro dos parâmetros estabelecidos.

Retorne um JSON em uma linha sem espaços em branco e com todas as chaves começando com a primeira letra maiúscula e o restante minúsculo, no valor das chaves mantenha o padrão: primeira letra da primeira palavra como maiúscula e restante minúsculo. Não crie uma chave em volta das

informações solicitadas.

3.5.4 Discussão sobre o prompt

3.5.4.1 Clareza e especificação do contexto

O prompt é introduzido com uma clara definição de contexto: “Você é um especialista em atendimento de suporte técnico em uma fintech”. Essa instrução estabelece uma expectativa de expertise no domínio específico, preparando o modelo para interpretar os dados de maneira contextualizada. A continuação, “Você receberá um JSON com o conteúdo de todo o atendimento realizado, podendo conter conversas de chat, whatsapp, email e outros”, delimita o escopo dos dados a serem analisados, garantindo que o modelo compreenda o formato e a abrangência das informações fornecidas.

3.5.4.2 Estrutura e organização

O prompt é organizado, detalhando as informações específicas a serem extraídas de cada interação. Por exemplo, para o campo “Motivo”, o prompt solicita que o modelo identifique o “Principal motivo pelo qual o cliente entrou em contato”, oferecendo sugestões como “Dúvidas, Problemas, Solicitações, Sugestões, Reclamações, Acompanhamento Interno”. Essa abordagem guiada minimiza interpretações errôneas e direciona o modelo a respostas dentro do escopo esperado.

A seção de *Tags* é abrangente, fornecendo uma lista extensa de termos como “Cartão de Crédito, Cartão de Débito, Cartão Pré-Pago, Bloqueio de Cartão, etc”. Essa especificidade não apenas facilita a categorização das interações, mas também permite a criação de novas *tags*, caso as pré-definidas não sejam adequadas: “Se não encontrar uma tag, pode criar novas tags”.

A lista de *tags* no *prompt* foi levantada a partir de uma semana de análises de atendimentos utilizando uma versão mais aberta para as *tags*.

3.5.4.3 Especificação do formato de saída

O formato de saída é especificado para assegurar a uniformidade dos dados: “Retorne um JSON em uma linha sem espaços em branco e com todas as chaves começando com a primeira letra maiúscula e o restante minúsculo”. Além disso, o *prompt* instrui que “no valor das chaves, mantenha o padrão: primeira letra da primeira palavra como maiúscula e restante minúsculo”. Essas diretrizes detalhadas são fundamentais para garantir a consistência no processamento automatizado e na análise posterior dos dados.

3.5.4.4 Normalização de respostas e limitação do escopo

Os campos “Motivo”, “Tags”, “Sentimento”, “SentimentoEmoji”, “SentimentoFinal”, “SentimentoFinalEmoji” e “Produtos” foram elaborados para garantir a normalização das respostas, o que é fundamental para a construção de relatórios quantitativos. A normalização desses campos permite a padronização das respostas, facilitando a agregação de dados e a realização de análises estatísticas.

Para os campos “Produtos” e “Tags”, o prompt instrui o modelo a priorizar a lista de referência, mas permite a criação de novos itens, se necessário. Para “Produtos”: “Se não encontrar um produto na lista, crie novos produtos”. E para “Tags”: “Se não encontrar uma tag, pode criar novas tags”. A criação de novas tags e produtos é permitida apenas quando os itens listados não se aplicam, assegurando que o modelo forneça respostas relevantes. Isso possibilita a identificação de novos produtos que os clientes podem estar demandando e de temas ainda não mapeados na lista de tags.

Para a categorização de sentimentos, é fornecida uma “LISTA DE SENTIMENTOS”, com exemplos como satisfeito (😊), neutro (😐), esperançoso (🙏), curioso (🤔), preocupado (😟), frustrado (😡), confuso (😕), desesperado (😱), indiferente (😒), irritado (😠), triste (😞), ansioso (😰). Essa lista é essencial para padronizar as respostas, evitando ambiguidades e promovendo consistência.

3.5.4.5 Verificação de respostas

A conclusão do *prompt* inclui uma instrução para verificação das respostas: “Verifique cuidadosamente suas respostas antes de concluir para assegurar que todas as informações fornecidas se alinhem estritamente com os requisitos”. Esta orientação ajudou a garantir que as respostas estejam corretas e aderentes às especificações fornecidas, mantendo a integridade dos dados. A atenção às “datas das mensagens trocadas” é enfatizada para assegurar a precisão temporal na análise.

3.5.4.6 Campos qualitativos

Além dos campos quantitativos normalizados, o prompt também inclui campos qualitativos que permitem uma análise contextualizada dos atendimentos ao cliente. Esses campos incluem “Assunto”, “Problema”, “Sugestao”, e “Resumo”, e são projetados para capturar nuances e detalhes que não podem ser facilmente quantificados.

3.5.4.7 Assunto

O campo “Assunto” solicita ao modelo que crie um título que sintetize o tema principal do atendimento. A instrução é: “Crie um assunto que ilustre da melhor forma possível este atendimento”. Este campo foi projetado pois nem sempre um assunto de chamado

de atendimento é totalmente compatível com o conteúdo, ajudando a identificar o foco principal de cada caso.

3.5.4.8 Sugestão

O campo “Sugestão” é projeto para utilização em melhoria contínua, solicitando que o modelo forneça sugestões de melhorias tanto para o analista de atendimento quanto para os processos internos da empresa. A instrução detalha: “Por favor, forneça sugestões de melhoria para o analista de atendimento e nossos processos. O que o analista poderia ter feito para melhorar nosso atendimento? Quais melhorias podemos implementar em nossos processos?” Além disso, o prompt pede sugestões de marketing e produto, destacando oportunidades para melhorar a experiência do cliente e explorar “gatilhos reptilianos” — elementos psicológicos que podem influenciar o comportamento do consumidor. Esse campo é essencial para identificar áreas de melhoria e inovação, bem como para reforçar práticas bem-sucedidas.

3.5.4.9 Resumo

Por fim, o campo “Resumo” solicita uma síntese concisa de todo o atendimento, incluindo informações como datas, números e nomes relevantes: “Resumo conciso do atendimento, interpretando todos os dados. Inclua detalhes como datas, números, nomes, etc”. Este campo serve para condensar a interação em um formato resumido, facilitando a revisão e análise de casos específicos.

3.5.5 Preparo dos dados

3.5.5.1 Recuperação dos dados

Como o atendimento é omnichannel e focado em canais de texto, foi necessário coletar dados de diferentes fontes, como WhatsApp, App, Chat web, Discord, e-mail e respostas no ambiente logado das contas dos clientes. Isso introduziu um risco de consistência de dados que impactou a qualidade dos resumos, pois a cronologia das conversas pode estar fragmentada entre essas bases de dados. Como este é um estudo de prova de conceito (PoC), reorganizar esses dados de forma cronológica não foi viável para a versão utilizada no estudo. No entanto, esse problema foi mitigado incluindo instruções diretamente no prompt, como mostrado a seguir:

VERIFIQUE AS DATAS ANTES DE ESCREVER PARA manter em ordem cronológica.

3.5.5.2 Limpeza dos dados

Remoção de caracteres não relevantes, como HTML, espaços em branco e quebras de linha duplicadas. Os dados de um atendimento (chats, e-mails, notas internas dos agentes) foram organizados em JSON.

A seguir, apresenta-se uma amostra de dados de um atendimento completo. Neste exemplo, as quatro primeiras mensagens resultam de um resumo gerado a partir de uma conversa no WhatsApp, utilizando o mesmo prompt. No JSON apresentado, todas as informações que poderiam identificar pessoas ou links foram substituídas por “[Cliente]”, “[Agente de atendimento]” e “[Link]”.

```

1  {
2    "messages": [
3      {
4        "content": {
5          "messagingSessionSummaries": [
6            {
7              "Summary": "Em 31 de julho de 2024, a cliente [Cliente]
                           entrou em contato via WhatsApp relatando problemas para
                           acessar sua conta e a necessidade de atualizar o endereço
                           para a entrega de um cartão. A atendente [Agente de
                           atendimento] orientou a cliente a redefinir a senha
                           através de um e-mail enviado e a atualizar o endereço via
                           ticket de suporte. A cliente foi instruída a seguir os
                           procedimentos e ficou esperançosa com a resolução.",
8              "StartTime": "2024-07-31T17:54:29.000Z",
9              "EndTime": "2024-07-31T18:38:03.000Z"
10             },
11             {
12              "Summary": "Em 1 de agosto de 2024, a cliente [Cliente]
                           entrou em contato via WhatsApp relatando que não
                           conseguia acessar sua conta após ter seu celular roubado
                           e perder o número de telefone cadastrado. O analista [
                           Agente de atendimento] verificou a situação e informou
                           que seria necessário um bloqueio temporário da conta e a
                           validação de identidade para resolver o problema. A
                           cliente forneceu os dados solicitados e foi informada que
                           um analista entraria em contato por telefone para
                           finalizar o processo. A cliente terminou o atendimento
                           esperançosa de que o problema seria resolvido em breve.",
13              "StartTime": "2024-08-01T18:46:03.000Z",
14              "EndTime": "2024-08-01T19:37:52.000Z"
15             },
16             {
17              "Summary": "Em 02 de agosto de 2024, a cliente [Cliente]
                           entrou em contato via WhatsApp relatando dificuldades

```



```

para acessar sua conta, pois o código de verificação
estava sendo enviado para um número de telefone antigo.
Após várias tentativas e frustrações, o analista [Agente
de atendimento] reenviou o e-mail com o procedimento
correto para alterar o número de telefone e a senha. A
cliente foi orientada a seguir os passos e aguardar a
confirmação por e-mail. O atendimento foi finalizado com
a cliente esperançosa de que o problema seria resolvido
.",
"StartTime": "2024-08-02T18:15:08.000Z",
"EndTime": "2024-08-02T18:52:53.000Z"
},
{
"Summary": "Em 05/08/2024, a cliente [Cliente] entrou em
contato via WhatsApp relatando problemas para acessar o
aplicativo devido à troca de número de telefone, o que
impedia a redefinição de senha. A analista [Agente de
atendimento] orientou a cliente a seguir os passos do
ticket de suporte já aberto e aguardar o retorno do time
responsável após o envio dos documentos necessários. A
cliente expressou frustração com a dificuldade de acesso,
mas a conversa terminou de forma neutra.",
"StartTime": "2024-08-05T17:17:59.000Z",
"EndTime": "2024-08-05T17:55:12.000Z"
}
],
"messagesIntranet": null,
"liveChatTranscriptSummaries": [],
"emailMessages": [
{
"Message": "Olá, [Cliente]. Estamos encaminhando abaixo o
procedimento a ser realizado para a alteração de seus
dados cadastrais. É muito importante que você nos informe
, respondendo a esse mesmo ticket, quando finalizar o
envio dos documentos solicitados, para que possamos
seguir com as alterações. Após recebidos os documentos
através do link e a informação de que o processo foi
concluído, o retorno será dado em até 01 (um) dia útil.
Para iniciar o processo de Alteração dos Dados Cadastrais
da sua Conta Digital, pedimos por gentileza, para
acessar o link abaixo, através do seu smartphone: Link: [
Link] Tenha em mãos o seu RG ou a sua CNH, para que
possamos validar a sua identidade e, ao final do processo
, informe os dados cadastrais que deseja alterar.
Lembramos que a Efí não solicita nenhum dado de acesso à
sua conta nem a realização de transação financeira por
redes sociais, WhatsApp, telefone, e-mail ou outros

```

```

32         canais. Conte com a gente! Atenciosamente, [Agente de
33         atendimento]",
34         "EntryDate": "2024-08-02T18:31:05.000Z",
35         "ActorName": "System"
36     },
37     {
38         "Message": "Olá, [Cliente]. Confirmando o recebimento da sua
39         comunicação e informo que o setor responsável foi
40         acionado. Por favor, aguarde o retorno do nosso setor,
41         que será conduzido por meio deste ticket. Permaneço à
42         disposição! Conte com a gente! Atenciosamente, [Agente de
43         atendimento]",
44         "EntryDate": "2024-08-01T20:09:14.000Z",
45         "ActorName": "System"
46     }
47 ]
48 }

```

Código 3 – Amostra de dataset

3.5.6 Hiperparâmetro temperatura

O hiperparâmetro “temperatura”¹ nas APIs da OpenAI desempenha um papel crucial na determinação da criatividade e variação nas respostas geradas pelos modelos de linguagem. Esse parâmetro controla a aleatoriedade das previsões: valores mais baixos (próximos de 0) fazem com que o modelo produza respostas mais determinísticas e previsíveis, enquanto valores mais altos (próximos de 1) permitem maior diversidade nas saídas, incentivando respostas mais criativas.

Para diferentes tipos de tarefas, recomenda-se ajustar a temperatura conforme a necessidade específica. Por exemplo, para tarefas de criação de conteúdo ou brainstorming, utilizam-se frequentemente valores de temperatura mais altos, como 0,7 ou 0,8, para promover a geração de ideias mais variadas e criativas. Em contraste, para tarefas que exigem maior precisão e consistência, como a avaliação de atendimentos ao cliente, é preferível utilizar valores baixos de temperatura para garantir que as respostas sejam mais objetivas e alinhadas com a realidade do contexto.

¹ OPENAI. API Reference - Temperature. Disponível em: <https://platform.openai.com/docs/guides/text-generation/how-should-i-set-the-temperature-parameter>. Acesso em: 09 ago. 2024.

Neste estudo, onde se busca um comportamento mais “conservador” e aderente aos dados, experimentaram-se diferentes configurações de temperatura e, após análise, optou-se por utilizar o valor de 0,1. Este ajuste foi escolhido para assegurar que o modelo operasse de maneira consistente e com foco na precisão, minimizando desvios criativos que poderiam comprometer a avaliação dos atendimentos.

3.5.7 Automação

O processo de avaliação é disparado automaticamente cada vez que um chamado é encerrado. Se o cliente ou os agentes reabrem o chamado, atualizam e encerram novamente, a avaliação é atualizada, conforme ilustrado na Figura 4 - Visão geral do processo de atendimento omnichannel e consumo da API.

3.5.8 gpt-4o-mini

Foram realizados testes utilizando o modelo gpt-4o-mini, que é mais rápido e apresenta um custo 97% menor em comparação ao gpt-4o. Mais detalhes no Capítulo 4 - Resultados.

4 RESULTADOS

4.1 ANONIMIZAÇÃO DOS CONTEÚDOS

Para assegurar a privacidade de todos os envolvidos e proteger dados sensíveis, foi realizada a anonimização dos conteúdos neste TCC conforme as seguintes diretrizes:

1. Os nomes dos clientes foram alterados para “[CLIENTE]” e os nomes dos especialistas de suporte da Efí Bank foram substituídos por “[ESPECIALISTA DE SUPORTE]”.
2. Qualquer link presente nas conversas foi anonimamente representado como “[LINK]”.
3. Números específicos de tickets foram substituídos pela expressão “Ticket Nº [TICKET]”.
4. Além disso, quaisquer outros dados sensíveis encontrados, como endereços de e-mail, números de telefone, números de cartões de crédito, endereços físicos, entre outros, foram substituídos pelo termo correspondente, utilizando o padrão “[TIPO_DE_DADO]” para designar o tipo de informação original. Por exemplo, números de telefone foram substituídos por “[TELEFONE]”.

4.2 ANONIMIZAÇÃO DOS GRÁFICOS DE RESULTADOS

Para garantir a confidencialidade dos dados estratégicos da Efí Bank, adotamos medidas de anonimização nos gráficos de resultados, conforme descrito a seguir:

1. Os nomes das *Tags* e dos produtos mencionados nos gráficos foram anonimizados.
2. Além disso, os valores apresentados nos gráficos foram normalizados.

Essas medidas foram implementadas para manter a integridade das análises apresentadas, ao mesmo tempo em que protegem os interesses estratégicos da Efí Bank.

4.3 AVALIAÇÕES

Entre os dias 15/04/2024 e 27/08/2024 foram realizadas 113.673 avaliações. Com base nos resultados, foi possível verificar que o modelo **gpt-4o** apresentou uma performance consistente em capturar nuances nas interações, identificar corretamente os problemas relatados pelos clientes e sugerir melhorias para o atendimento e produtos. Contudo, a sensibilidade ao prompt e o fenômeno de alucinações ainda se mostraram como desafios significativos, necessitando ajustes constantes no prompt para melhorar a coerência e a precisão das respostas.

4.3.1 Amostra selecionada

```

1  {
2    "messages": [
3      {
4        "content": {
5          "messagesIntranet": [
6            {
7              "Message": "Olá, [CLIENTE]! Queremos oferecer a você um
                           serviço seguro e transparente, por isso precisamos de sua
                           colaboração. A Efí enquanto Instituição de Pagamento,
                           possui processos internos que visam garantir a qualidade
                           e a segurança das transações, bem como o cumprimento dos
                           nossos Termos e Condições de Uso (contrato) e Exigências
                           Regulatórias. Para isso, contamos com um sistema de
                           monitoramento que nos alerta quando é feita alguma ação
                           ou interação com sua conta que necessite de análise. Em
                           uma dessas verificações, tivemos apontamentos que nos
                           levaram a suspender preventivamente as operações na sua
                           conta digital. Para darmos continuidade ao processo de
                           análise, solicitamos que nos envie algumas informações
                           adicionais: - Descreva com suas palavras a finalidade da
                           utilização da conta digital. - Encaminhe documentos e
                           informações que comprovem sua atividade comercial e/ou
                           profissional. Para verificar quais documentos são aceitos
                           , clique [LINK]. - Caso não tenha nenhum dos documentos
                           indicados no link acima, nos informe como é firmada a
                           relação de prestação de serviços entre as partes (você e
                           seu cliente). Para maiores informações sobre os motivos
                           que levaram a suspensão das operações na sua conta,
                           acesse: [LINK] Importante! A tratativa desta análise será
                           feita exclusivamente por meio deste ticket. O prazo para
                           resposta é de até 05 (cinco) dias úteis a cada interação
                           . Caso não tenhamos retorno em até 30 dias corridos, sua
                           conta digital poderá ser encerrada em definitivo. Conte
                           com a gente! Efí S.A.",
8              "EntryDate": "2024-06-07T09:40:03.000Z",
9              "ActorName": "System"
10             }, {
11              "Message": "Olá, Agradeço a atenção. Muito obrigado pela
                           cordialidade. Segue acima o contrato de locação em meu
                           nome, onde consta a matriz do meu empreendimento. Para
                           agregar a este ticket, passo a seguinte informação sobre
                           a funcionalidade do meu empreendimento: Funciona 24 horas
                           , sendo que o que mais se destaca é o delivery, tanto de
                           comida quanto de bebidas alcoólicas em geral. Recebemos

```

```

dos nossos clientes entre 68% e 80% do total da operação
via Pix, realizado diretamente pelo WhatsApp com a
plataforma integrada. Não podemos parar a operação, ainda
mais considerando que esse valor não representa 5% de um
dia, pois temos 3 pontos de venda. Toda vez que ocorre
um problema com a conta na qual a chave Pix está
cadastrada no aplicativo de pedidos via WhatsApp, tenho
um grande transtorno, pois preciso entrar em contato com
a empresa do sistema ou avisar cliente por cliente,
fornecendo outra chave Pix. Por favor, peço que retorne o
acesso à minha conta o quanto antes. Escolhi sua empresa
porque ouvi falar muito bem dela. Atenciosamente, [
    CLIENTE]",
    "EntryDate": "2024-06-09T22:18:38.000Z",
    "ActorName": "EndUser"
}, {
    "Message": "Eu utilizo a conta digital para o recebimento e
pagamentos por meio de Pix. Presto serviço de marketing,
aumentando o número de clientes de outras empresas
digitais. Trabalho com indicação de clientes para Casas
de aposta, Hotmart, Monetizze, entre outros. Recebo
comissão pelo fato de o cliente comprar ou assinar.
Quando o cliente faz assinatura de um produto, geralmente
cursos, recebo comissão recorrente; e, quando o produto
é de venda única, recebo a comissão única. Eu capto
clientes e ganho comissão.",
    "EntryDate": "2024-06-28T01:48:17.000Z",
    "ActorName": "EndUser"
}, {
    "Message": "Possuo um instagram [LINK]. a relação contratual
entre minha empresa e os clientes é estabelecida por meio
dos termos e condições das plataformas Monetizze,
Hotmart, amazon, mercadolibre e etc. A comprovação da
prestação de serviços ou entrega de produtos pode ser
feita através de registros de vendas, emissão de notas
fiscais, relatórios detalhados e suporte ao cliente
oferecido pelas plataformas.",
    "EntryDate": "2024-07-18T18:54:18.000Z",
    "ActorName": "EndUser"
}, {
    "Message": "Parece que você não entendeu. Eu não vendo
produto, eu ganho comissão na divulgação apenas.",
    "EntryDate": "2024-07-23T19:58:58.000Z",
    "ActorName": "EndUser"
}
],
"liveChatTranscriptSummaries": [],

```

```

29     "emailMessages": [
30         {
31             "Message": "Olá, O meu banco não consta nas opções da página
                        [LINK], eu fui orientado pelo suporte a enviar por aqui
                        os dados. segue anexo minha conversa com o suporte.",
32             "EntryDate": "2024-08-05T18:14:19.000Z",
33             "ActorName": "EndUser"
34         }, {
35             "Message": "Olá, [CLIENTE]. Após o encerramento da conta,
                        eventuais recursos disponíveis como saldo, poderão ser
                        transferidos somente via TED, para contas bancárias de
                        mesma titularidade. Essa ação deverá ser executada via [
                        LINK]. Conte com a gente! Atenciosamente, Departamento
                        Antifraude - Efí S",
36             "EntryDate": "2024-08-05T17:56:29.000Z",
37             "ActorName": "System"
38         }, {
39             "Message": "Olá, segue os dados do a transferência do saldo
                        total abaixo: Banco: CloudWalk Instituição de Pagamento
                        Agência (sem dígitos): [TELEFONE] Conta corrente com
                        dígito: [TELEFONE]",
40             "EntryDate": "2024-08-02T23:15:38.000Z",
41             "ActorName": "EndUser"
42         }, {
43             "Message": "Olá, [CLIENTE]. Após análise, decidimos pelo
                        encerramento definitivo da sua conta digital. Este
                        procedimento visa alinhar os perfis dos clientes com
                        nossas políticas internas e comerciais, legitimando as
                        operações da Efí dentro das determinações dos órgãos
                        reguladores e adequando-as aos termos de uso. Reforçamos
                        que a análise contínua de contas é um processo
                        obrigatório para as Instituições de Pagamentos e visa a
                        segurança tanto do Cliente quanto da Efí. Após o
                        encerramento da conta, eventuais recursos disponíveis
                        como saldo, poderão ser transferidos somente via TED,
                        para contas bancárias de mesma titularidade. Essa ação
                        deverá ser executada via [LINK]. Cientes do seu
                        conhecimento acerca do que estabelece nossos Termos e
                        Condições de Uso, também denominado Contrato, contamos
                        com a sua compreensão e permanecemos à disposição.
                        Atenciosamente, Efí S",
44             "EntryDate": "2024-07-30T00:27:37.000Z",
45             "ActorName": "System"
46         }, {
47             "Message": "Olá, [CLIENTE]. Agradecemos pelo retorno.
                        Gentileza nos encaminhe uma das notas fiscais mencionadas
                        para verificarmos. Atenciosamente, Efí S",

```

```

48     "EntryDate": "2024-07-23T11:54:43.000Z",
49     "ActorName": "System"
50 }, {
51     "Message": "Olá, [CLIENTE]. Agradecemos o retorno. Destacamos
        que sua conta está atualmente em processo de análise, um
        procedimento padrão necessário para cumprir com as
        Exigências Regulatórias e os Termos e Condições de Uso do
        nosso contrato. Gostaríamos de salientar que a
        finalidade desta solicitação é compreender melhor o seu
        empreendimento e validar as transações efetuadas por meio
        da plataforma Efí. Portanto, solicitamos que: Confirme
        se possui um website e/ou redes sociais para fins
        comerciais. Explique como é estabelecida a relação
        contratual de prestação de serviço entre sua empresa e
        seus clientes. Em caso de disputas sobre valores,
        gostaríamos de saber como é possível comprovar que o
        serviço foi prestado ou que o produto foi entregue. Esta
        informação nos ajudará a entender melhor seu negócio e
        garantir transações seguras e confiáveis por meio da Efí.
        Atenciosamente, Efí S",
52     "EntryDate": "2024-07-06T00:11:13.000Z",
53     "ActorName": "System"
54 }, {
55     "Message": "Olá, [CLIENTE]. Agradecemos pelo retorno.
        Informamos que não foi possível validar a prestação de
        serviço através dos documentos apresentados. Conforme rol
        de documentos aceitos informado anteriormente, o
        contrato de prestação de locação é válido, desde que a
        atividade exercida seja locações de imóveis. Aguardamos
        seu retorno. Atenciosamente, Efí S",
56     "EntryDate": "2024-06-17T23:13:54.000Z",
57     "ActorName": "System"
58     }
59 ]
60 }
61 }
62 ]
63 }

```

Código 4 – Amostra selecionada

4.3.2 Resultado da avaliação

Tabela 2 Resultado da avaliação

Assunto	Suspensão e encerramento de conta digital por análise de conformidade
Motivo	Problemas
Problema	Conta digital suspensa e posteriormente encerrada devido a análise de conformidade e falta de validação de documentos.
Produtos	Conta digital, Pix
Resolução	Conta encerrada definitivamente após análise. Saldo disponível será transferido via TED para conta bancária de mesma titularidade.
Resumo	Em 07 de junho de 2024, a conta digital de [CLIENTE] foi suspensa para análise de conformidade. [CLIENTE] forneceu informações sobre seu empreendimento e documentos, mas a validação não foi concluída com sucesso. Em 30 de julho de 2024, a conta foi encerrada definitivamente. [CLIENTE] foi informado que o saldo disponível seria transferido via TED para uma conta bancária de mesma titularidade. O cliente expressou preocupação e frustração durante o processo, mas o sentimento final foi neutro.
Sugestão	O analista poderia ter sido mais claro e específico sobre os documentos necessários desde o início, evitando múltiplas interações. Implementar um sistema de upload de documentos diretamente no portal do cliente pode agilizar o processo. Para o marketing, explorar a segurança e conformidade como pontos fortes pode atrair clientes que valorizam essas características. Para o time de produto, considerar a criação de uma seção no aplicativo onde os clientes possam acompanhar o status da análise de documentos em tempo real e receber notificações sobre pendências.
Tags	Bloqueio de conta, encerramento de conta, documentação pendente, validação de documentos, chave Pix, transferência Pix

4.3.3 Considerações sobre a avaliação

A avaliação está alinhada com os eventos ocorridos durante o atendimento. Para aprimorar o detalhamento e a padronização das datas, foram realizados ajustes no prompt

nas versões mais recentes do projeto, desenvolvidas após este estudo:

“Você receberá datas no formato “2024-06-07T09:40:03.000Z” mas para as respostas que envolverem datas, adote o formato dd/mm/aaaa.”

(...)

“Resumo: Resumo do atendimento. Inclua detalhes como datas, números, etc. Mencione os nomes dos envolvidos. VERIFIQUE AS DATAS ANTES DE ESCREVER PARA manter em ordem cronológica. IGNORE O PARÂMETRO “ActorName”. Você não deve, em hipótese alguma, usar termos como “o sistema respondeu” e sim o nome da pessoa que assina a mensagem. Não omita detalhes importantes.”

Após os ajustes no prompt, o resumo mudou para:

Resumo	Em 07/06/2024, a Efí S.A. [CLIENTE] sobre a suspensão preventiva de sua conta digital, solicitando informações adicionais e documentos para análise. [CLIENTE] respondeu em 09/06/2024, fornecendo detalhes sobre seu empreendimento e solicitando a reativação da conta. Em 28/06/2024 e 18/07/2024, [CLIENTE] enviou mais informações sobre suas atividades comerciais. Em 23/07/2024, [CLIENTE] esclareceu que não vende produtos, mas ganha comissões por divulgação. Em 30/07/2024, a Efí S.A. informou o encerramento definitivo da conta de [CLIENTE]. Em 02/08/2024, [CLIENTE] forneceu dados bancários para transferência do saldo. Em 05/08/2024, a Efí S.A. orientou sobre a transferência via TED. Em 19/08/2024, [CLIENTE] relatou erro na transferência, e a Efí S.A. informou que a solicitação foi encaminhada para o departamento responsável.
--------	---

4.4 DASHBOARD COM RESULTADOS QUANTITATIVOS

A análise quantitativa dos resultados foi conduzida por meio de um *dashboard* (ver Figura 5), que permite o monitoramento em tempo real do desempenho dos atendimentos e das avaliações geradas pelos modelos.

O *dashboard* se destaca especialmente na identificação dos problemas e produtos mais recorrentes em diferentes intervalos de tempo, com ênfase no fator tempo real.

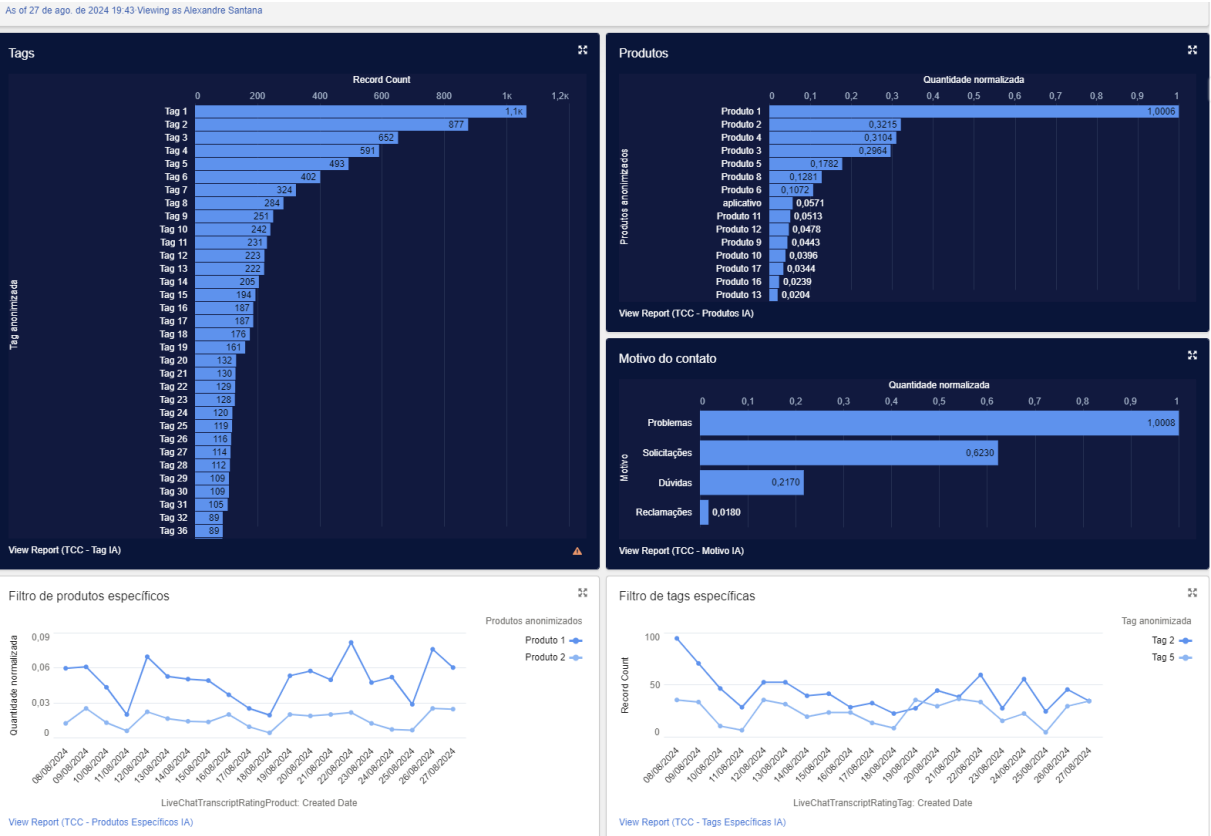


Figura 5 – Visão geral do dashboard com análise quantitativa de tags, produtos, motivos de contato e filtros específicos, permitindo o monitoramento detalhado do desempenho dos atendimentos e a identificação de padrões relevantes.

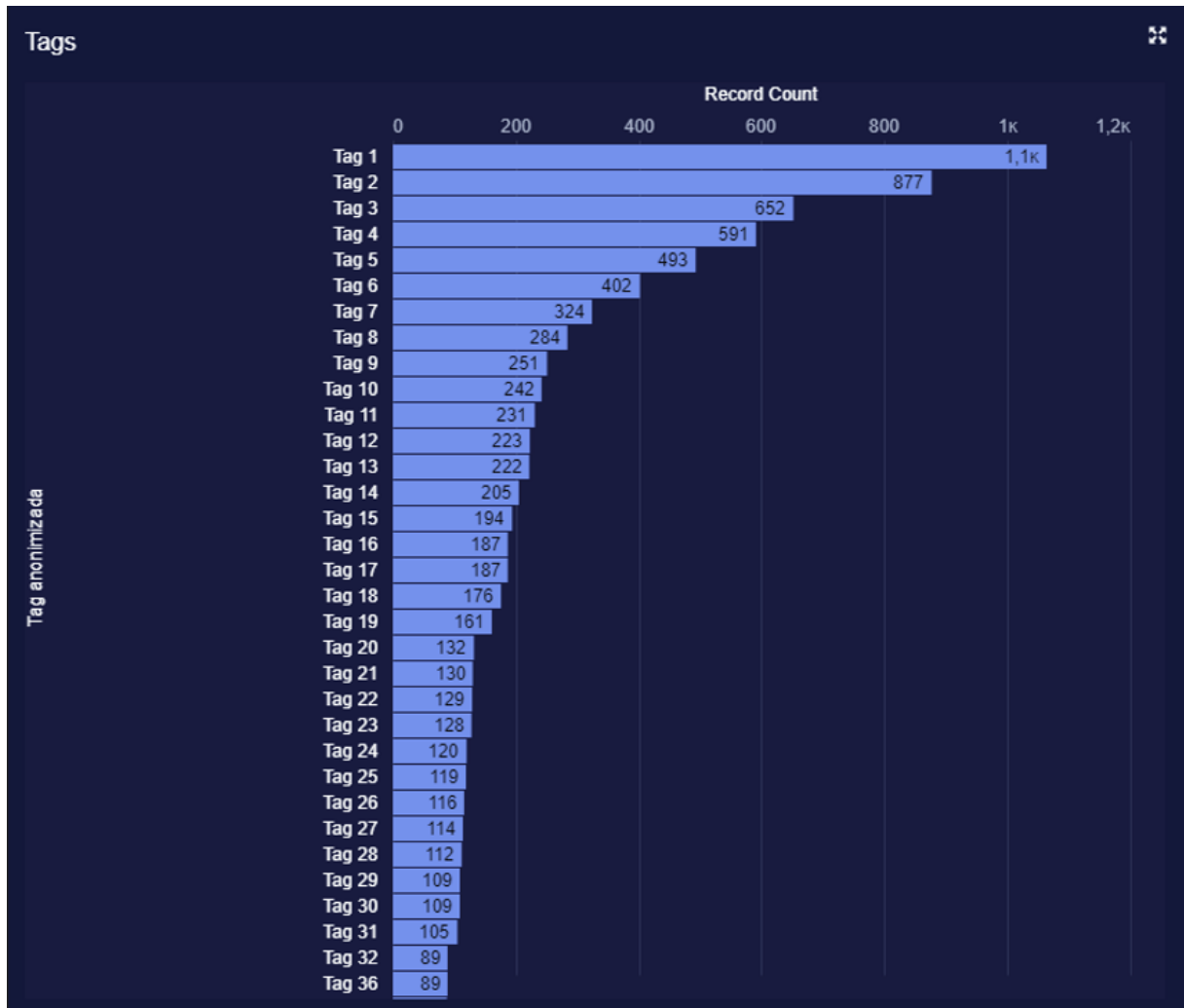


Figura 6 – Distribuição das tags mais utilizadas em atendimentos, destacando a frequência com que cada tag foi aplicada, com “Tag 1” sendo a mais comum.



Figura 7 – Quantidade normalizada de menções a produtos durante os atendimentos, com “Produto 1” sendo o mais referenciado em relação aos demais.



Figura 8 – Motivos dos contatos classificados em problemas, solicitações, dúvidas e reclamações, com “Problemas” sendo o motivo mais frequente.

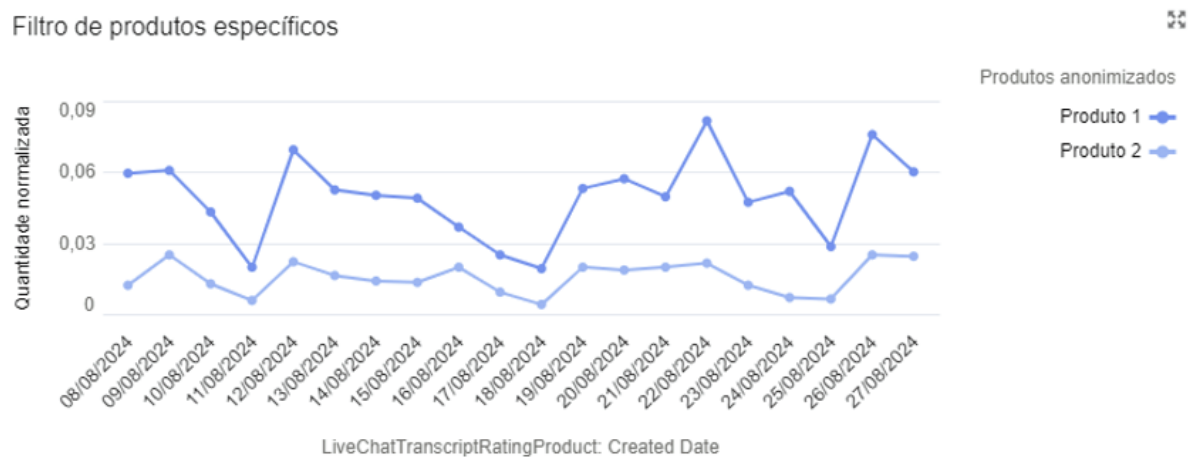


Figura 9 – Evolução temporal da quantidade normalizada de dois produtos anonimizados, analisando como a menção a esses produtos variou ao longo do tempo.

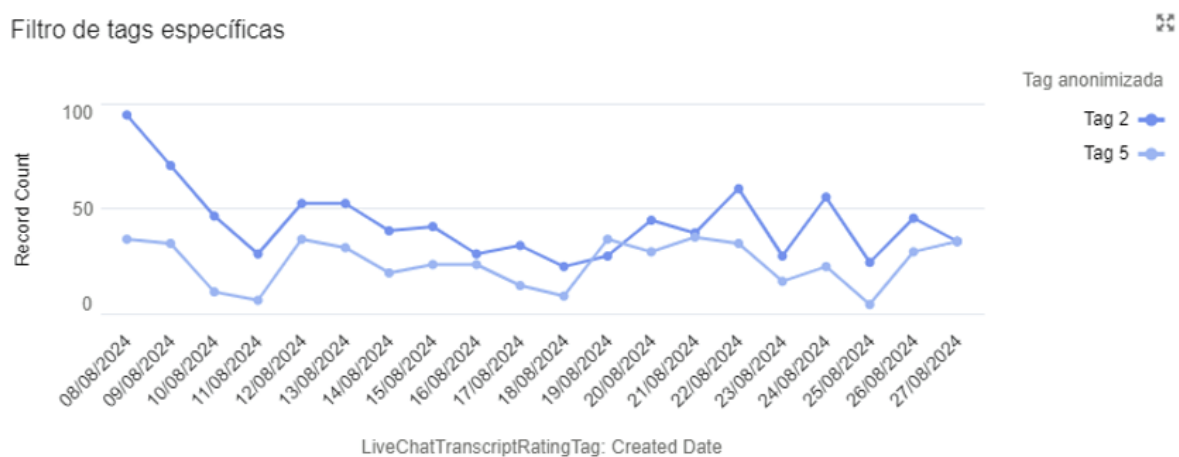


Figura 10 – Evolução temporal do uso de duas tags anonimizadas, destacando as variações na frequência dessas tags ao longo do tempo.

4.5 PRINCIPAIS DESAFIOS

Após todo o processo de desenvolvimento e implantação, foram elencados os seguintes desafios como os mais relevantes:

- Sensibilidade ao prompt
- Alucinações
- Custos
- Volume de dados
- Normalização de dados quantitativos

4.5.1 Sensibilidade ao prompt e alucinações

Os modelos de grande linguagem não realizam raciocínio no sentido humano. Eles funcionam como Processamento de Linguagem Natural (PLN) avançado, processando o input e organizando a informação com base em elementos e aprendizados de seu vasto conjunto de dados de treinamento. Por essa razão, o resultado gerado é altamente sensível à estrutura do prompt utilizado.

A construção de um prompt ainda depende muito de tentativa e erro. Na versão de prompt deste projeto (3.5.3), foram realizados dezenas de testes (provavelmente mais de 100, embora não tenha sido contabilizado precisamente).

4.5.2 Alucinações

Quando o prompt instrui de forma imperativa que o modelo organize informações que não estão presentes no dataset, é quase certo que ele inventará dados para atender à solicitação. Neste contexto, vale mencionar três exemplos:

- **Análise de sentimento:** É comum que, em alguns atendimentos, o cliente não expresse agradecimento explícito quando sua dúvida é resolvida. Nesses casos, pode ser difícil determinar se o cliente ficou satisfeito ou qual foi seu sentimento. Mesmo assim, o modelo tende a inferir sentimentos. Em uma versão recente do prompt, isso foi mitigado com o seguinte ajuste:

“Sentimento: Responda em uma palavra qual o sentimento do cliente. Selecione da LISTA DE SENTIMENTOS, informada acima. NÃO TIRE CONCLUSÕES PRECIPITADAS. Na dúvida, responda apenas “indefinido”.”

- **Datas e horários:** Devido a um bug, inicialmente não estávamos enviando as datas das interações, o que resultou em resumos com datas inventadas pelo modelo. Após corrigirmos o dataset, incluindo as datas corretas, o modelo passou a gerar resumos com datas precisas.
- **Contextos de atendimento com pouca informação:** Mesmo em atendimentos com pouco conteúdo trocado entre cliente e agente, o modelo pode criar uma narrativa completa a partir de poucas frases.

Esses desafios podem ser solucionados (ou mitigados) com o uso de prompts dinâmicos, ajustados conforme o volume e a natureza das interações (utilizando algoritmos para verificar a quantidade de caracteres ou interações), ou até mesmo utilizando diferentes agentes de IA que classificam a robustez ou a natureza do dataset, selecionando dinamicamente o prompt mais adequado.

4.5.3 Custos

O custo médio por avaliação utilizando o modelo gpt-4o é de \$0,01355. Considerando um volume de 100 mil avaliações mensais, o custo total seria de \$1.355,00 por mês. Com a cotação do dólar em R\$5,47, conforme verificado em 20 de agosto de 2024, o custo mensal em reais seria de R\$7.302,45.

Também foram realizados testes com o modelo GPT-4o-mini, que, embora tenha um custo aproximadamente 97% menor por 1 milhão de tokens em comparação ao gpt-4o, apresentou resultados abaixo do esperado em alguns campos de resposta, especialmente no resumo, sem que fossem feitos ajustes significativos na arquitetura do projeto.

4.5.4 Volume de dados

Um dos desafios foi que o prompt de entrada, para atendimentos longos envolvendo múltiplos canais como chats, WhatsApp e e-mails, frequentemente ultrapassava o limite da API. Para contornar isso, optamos por avaliar separadamente as conversas de chats e WhatsApp, integrando esses resultados à análise final do atendimento, que incluía os outros canais. Isso garantiu a eficiência da análise, apesar das limitações técnicas.

4.5.5 Normalização de dados quantitativos

Para definir uma lista eficaz de tags a ser incluída no prompt e garantir a normalização dos resultados, foi adotada a estratégia de deixar o algoritmo rodando por algumas semanas sem uma lista predefinida de tags. Durante esse período, o modelo gerou tags de forma autônoma, refletindo os tópicos mais recorrentes nos atendimentos. Em seguida, foi utilizado o modelo GPT-4o para, a partir dessas tags, construir uma lista abrangente dos principais tópicos. Com essa lista definida, o prompt foi ajustado para incluir as

tags predefinidas, instruindo o modelo a selecionar as mais relevantes, o que permitiu a normalização dos resultados. A seguir, apresenta-se o trecho do prompt ajustado com a lista de tags resultante dessa técnica.

Tags: Lista de principais palavras-chave do atendimento (Evite tags não relevantes como “atendimento ao cliente” ou “ticket de suporte”). Para reduzir a redundância entre esta e futuras avaliações, viabilizar normalização em relatórios e manter a informação relevante, aqui estão as tags principais: Cartão de Crédito, Cartão de Débito, Cartão Pré-Pago, Cartão Virtual, Emissão de Cartão, Bloqueio de Cartão, Cancelamento de Cartão, Perda/Roubo de Cartão, Segunda Via de Cartão, Transferência Pix, Cobrança Pix, Chave Pix, QR Code Pix, Erro Pix, Bloqueio de Pix, Devolução Pix, Abertura de Conta, Encerramento de Conta, Bloqueio de Conta, Conta Digital, Conta Física, Conta Jurídica, Conta Poupança, Redefinição de Senha, Troca de Senha, Bloqueio de Senha, Senha Eletrônica, Envio de Documentos, Documentos Pessoais, Documentos Empresariais, Documentação Pendente, Validação de Documentos, Contestação de Transação, Transação Não Reconhecida, Transação Bloqueada, Erro de Transação, Pagamento Pix, Pagamento Boleto, Pagamento de Contas, Pagamento Não Identificado, Pagamento Duplicado, Erro no Sistema, Erro no Aplicativo, Erro de Autenticação, Erro de Pagamento, Erro de Cadastro, Autenticação de Dispositivo, Autenticação em Dois Fatores, Validação de Identidade, Validação de Selfie, Autenticação Facial, Emissão de Boleto, Pagamento de Boleto, Boleto Não Registrado, Boleto Duplicado, Boleto Não Pago, Limites (para lidar com solicitações de aumento ou redução de limite), Estorno de Pagamento (para lidar com devoluções e reembolsos), Acesso ao Aplicativo (para problemas específicos de login ao aplicativo), Acesso Web (para problemas específicos de login a conta pelo navegador), Atualização de Cadastro (para mudanças em informações pessoais ou empresariais). Se não encontrar uma tag, pode criar novas tags.

5 CONCLUSÃO

O objetivo principal deste projeto, que era aplicar o modelo LLM GPT-4 na automação da avaliação de atendimento ao cliente e validar sua eficácia por meio de uma prova de conceito em ambiente real, foi plenamente atingido.

5.1 HIPÓTESE INICIAL E RESULTADOS

A hipótese inicial sugeria que o LLM poderia abstrair automações complexas de avaliação de atendimento, eliminando a necessidade de agentes investirem tempo na classificação de todos os chamados ou na construção e manutenção de formulários e fluxos necessários para realizar a tarefa manual de análise. Com isso, seria possível criar um sistema de avaliação flexível a longo prazo, que demandasse apenas ajustes em prompts para atender novas necessidades e melhorias, podendo ser realizado por profissionais não necessariamente de TI, como lideranças de atendimento ou analistas de qualidade.

Os resultados obtidos corroboram essa hipótese. A flexibilidade do LLM demonstrou ser um dos seus maiores pontos fortes, permitindo que a automação da avaliação fosse adaptada com facilidade para novas demandas e contextos, apenas ajustando os prompts. Esse processo simplificado permitiu que profissionais com conhecimento do processo de atendimento, mas sem expertise técnica, pudessem realizar esses ajustes, confirmando que o modelo atende à necessidade de flexibilidade e adaptação ao longo do tempo, sem a necessidade de envolvimento constante de profissionais de TI.

5.2 DISCUSSÃO DOS OBJETIVOS ESPECÍFICOS

5.2.1 Explorar e definir conceitos fundamentais de LLMs e engenharia de prompt

A engenharia de prompt se mostrou essencial para obter resultados satisfatórios. A capacidade de ajustar os prompts adequadamente determinou a qualidade das respostas geradas pelo modelo, evidenciando a importância do domínio técnico sobre essa área.

5.2.2 Avaliar a eficácia e a precisão do LLM comparando seus resultados com análises humanas

As análises geradas pelo modelo, embora não superem aquelas realizadas pelos melhores profissionais, demonstram ser superiores às executadas por analistas típicos. É importante ressaltar que a comparação relevante para essa tarefa não deve ser feita entre o desempenho do modelo de linguagem e o dos profissionais mais qualificados, mas

sim entre o modelo e o profissional que, usualmente, seria responsável por executar essa função.

Considerando que a execução dessas tarefas pode ser massante e que um único analista dificilmente consegue realizar todas as análises necessárias, limitando-se apenas a uma amostragem pequena, a automação se apresenta como uma solução eficaz. Isso é especialmente relevante diante do fato de que o algoritmo realiza dezenas de milhares de análises mensais a um custo aproximado ao custo médio de um analista de qualidade.

5.2.3 Identificar vantagens, desafios e oportunidades na automação cognitiva com LLMs

Durante a implementação, foram identificados desafios, como o processamento de grandes volumes de dados e a adaptação dos modelos de linguagem para cenários complexos. Um dos principais desafios foi o tamanho dos prompts de entrada, que, em atendimentos longos e multicanais, frequentemente excediam os limites suportados pela API. Isso exigiu uma análise segmentada das conversas de diferentes canais, como chats e WhatsApp, com a posterior integração dos resultados em uma análise unificada.

Outro desafio relevante foi o custo operacional. O uso de modelos menores, como o gpt-4o-mini, apresentou uma redução de 97% nos custos; no entanto, devido à complexidade da tarefa, as análises resultaram em menor detalhamento e precisão. Para viabilizar economicamente o uso de modelos mais baratos, considerou-se a simplificação das tarefas ou a adoção de estratégias, como prompts dinâmicos baseados na complexidade e no tamanho dos dados, além de análises segmentadas para cada aspecto do atendimento.

Apesar dos desafios, foram identificadas oportunidades significativas na automação com LLMs. A flexibilidade desses modelos permite automatizar tarefas que antes exigiriam grandes investimentos em software ou trabalho humano intensivo. A possibilidade de ajustar rapidamente os prompts, sem a necessidade de profissionais de programação ou modificações em códigos complexos, possibilita que líderes de atendimento ou analistas de qualidade realizem otimizações em tempo real. A avaliação de custos deve considerar não apenas o processamento do modelo, mas também o tempo de desenvolvimento, manutenibilidade e o custo de profissionais especializados.

Foi identificada a possibilidade de aplicar essa tecnologia em outras áreas, como na avaliação de processos comerciais e no acionamento de gatilhos automáticos para campanhas de marketing, além de outros processos paralelos, como detecção de fraudes e geração de insights para ajustes ou desenvolvimento de novos produtos. Dessa forma, o projeto não apenas otimiza a avaliação de atendimentos, mas também apresenta o potencial de transformar outros processos operacionais e estratégicos dentro da organização.

Por fim, foi comprovado que é viável, tanto em termos de custos quanto de qualidade, aplicar IAs generativas na automação da avaliação de atendimento ao cliente. O projeto encontra-se atualmente em produção e tem sido utilizado com sucesso pelos times de quali-

dade do atendimento. O feedback recebido até o momento tem sido positivo, confirmando a viabilidade e a eficácia dessa aplicação na automação da avaliação de atendimento ao cliente.

REFERÊNCIAS

BROWN, T. et al. Language models are few-shot learners. **Advances in Neural Information Processing Systems**, 2020.

CHANG, E. Y. **Examining GPT-4: Capabilities, Implications and Future Directions**. [S.l.], 2023. Disponível em: <https://www.researchgate.net/publication/374753069>. Acesso em: 24 ago. 2024. Disponível em: <https://www.researchgate.net/publication/374753069>.

CHOLLET, F. On the measure of intelligence. **arXiv preprint arXiv:1911.01547**, 2019.

LEGG, S.; HUTTER, M. **A Collection of Definitions of Intelligence**. [S.l.], 2007.

OPENAI. **Prompt Engineering Guide**. 2024. <https://platform.openai.com/docs/guides/prompt-engineering/strategy-test-changes-systematically>. Acesso em: 26 ago. 2024.

QIAO, S. et al. **Reasoning with Language Model Prompting: A Survey**. [S.l.], 2023. Disponível em: <https://arxiv.org/abs/2212.09597v8>. Acesso em: 24 ago. 2024. Disponível em: <https://arxiv.org/abs/2212.09597v8>.

ROSSETTI, R.; GARCIA, K. Inteligência artificial generativa: questões jurídicas e éticas em torno do chatgpt. **VirtuaJus**, v. 8, n. 15, p. 253–264, 2023.