

LUCAS ALENCAR BARROS

Redesenho de Processos em E-commerce com Agentes de IA: Um Estudo de
Caso Integrando BPM, VTEX e Correios

PROJETO DE GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO
APRESENTADO AO DEPARTAMENTO DE ENGENHARIA INDUSTRIAL
DA PUC-RIO, COMO PARTE DOS REQUISITOS PARA OBTENÇÃO
DO TÍTULO DE ENGENHEIRO DE PRODUÇÃO

Orientador: Gustavo Costa Gomes Moreira

Departamento de Engenharia Industrial
Rio de Janeiro, 24 de novembro de 2025.

Agradecimentos:

Agradeço, em primeiro lugar, à PUC-Rio, instituição onde tive a oportunidade de construir minha formação acadêmica e profissional, e que ofereceu o ambiente intelectual necessário para o desenvolvimento deste trabalho.

Ao meu orientador, Professor Gustavo Moreira, registro meu sincero reconhecimento pela orientação inteligente, pela clareza conceitual e pelo entusiasmo transmitido ao longo das disciplinas e reuniões deste projeto. Suas aulas e provocações intelectuais tiveram papel determinante na forma como compreendi a Engenharia de Produção e influenciaram diretamente a abordagem metodológica adotada nesta pesquisa.

À Brofitwear, agradeço pela abertura, confiança e colaboração durante o estudo de caso, fornecendo dados, contexto e espaço para testar soluções em um ambiente real de operações. Esta pesquisa só foi possível graças ao apoio e disponibilidade da equipe.

Aos meus familiares, agradeço pela base incondicional que sempre me permitiu perseguir desafios com segurança e dedicação. Em especial, agradeço à minha parceira, Laura Veiga, pelo apoio constante, pelas conversas que me mantiveram firme nos momentos de maior carga e pela paciência generosa ao longo de todo o período de escrita deste trabalho.

Por fim, agradeço a todos os colegas, amigos e profissionais que, de diferentes formas, contribuíram com discussões, testes, leituras e incentivos. Cada contribuição ajudou a tornar este projeto mais sólido e significativo.

Resumo

Este trabalho investiga a adoção de Agentes de Inteligência Artificial (IA) como instrumento de redesenho de processos em uma operação de e-commerce de médio porte, tomando como estudo de caso a Brofitwear. A pesquisa parte da perspectiva da Gestão por Processos de Negócio (BPM), compreendendo a organização como um encadeamento de atividades que transformam insumos em valor, no qual eficiência, qualidade e experiência do cliente dependem da integração de fluxos ponta a ponta.

A partir de um mapeamento e diagnóstico de processos no setor de marketing e atendimento, se identificou a necessidade de redesenhar a jornada do consumidor, do pré-venda ao pós-venda, em função de gargalos relacionados à comunicação, ruptura de informações e retrabalho entre canais. Para endereçar essas ineficiências, foi concebido um agente conversacional inteligente (*chatbot*) capaz de facilitar e automatizar interações críticas: recomendações de produtos, esclarecimento de políticas e dúvidas, rastreamento de pedidos (WISMO) e apoio à logística reversa.

Diferente de *chatbots* determinísticos, o agente proposto integra a técnica de *Retrieval-augmented generation* (RAG), com acesso controlado a sistemas transacionais (CRM/OMS/WMS), políticas de governança, mecanismos de auditoria e o princípio de humano-no-circuito em situações sensíveis. Essa arquitetura garante rastreabilidade, conformidade e aprendizado incremental em um ambiente operacional real.

A contribuição esperada é dupla: prática, ao oferecer um roteiro replicável para criação, implantação e governança de agentes de IA integrados a processos e sistemas corporativos; e acadêmica, ao consolidar evidências sobre as condições de sucesso, riscos e trade-offs dessa tecnologia à luz dos princípios da Indústria 5.0: Cooperação homem-máquina, personalização, sustentabilidade e resiliência.

Palavras Chave:

Agente de Inteligência Artificial (IA). Comércio eletrônico (e-commerce). Gestão por Processos de Negócio (BPM). BPMS. Indústria 5.0. Mapeamento AS-IS/TO-BE. Experiência do cliente (CX). Eficiência operacional. VTEX. CORREIOS.

ABSTRACT:

This study investigates the adoption of Artificial Intelligence (AI) Agents as a means to redesign business processes within a medium-sized e-commerce operation, using Brofitwear as a case study. Grounded in the principles of Business Process Management (BPM), the organization is understood as a chain of activities that transform inputs into value, where efficiency, quality, and customer experience depend on well-integrated, end-to-end flows.

Following a process mapping and diagnosis in the marketing and customer service departments, the need to redesign the consumer journey, from pre-sale to post-sale, was identified due to issues of fragmented communication, inconsistent data, and rework across channels. To address these inefficiencies, an intelligent conversational agent (chatbot) was designed to facilitate and automate critical interactions, including product recommendations, policy inquiries, order tracking (WISMO), and reverse logistics support.

Unlike deterministic chatbots, the proposed agent integrates the Retrieval-Augmented Generation (RAG) system with controlled access to transactional systems (CRM/OMS/WMS), governance policies, audit mechanisms, and the human-in-the-loop principle for sensitive decisions. This architecture ensures traceability, compliance, and continuous learning within a live operational context.

The study's contribution is twofold: practical, by offering a replicable framework for the creation, deployment, and governance of process-integrated AI agents; and academic, by providing empirical evidence on the success factors, risks, and trade-offs of this technology in light of Industry 5.0 principles-human-machine cooperation, personalization, sustainability, and resilience

SUMÁRIO

1. Introdução:	1
2. Agentes de IA, RAG, Ferramentas e suas Aplicações:	4
2.1 Inteligência Artificial e Agentes Inteligentes	4
2.2 Arquitetura de Agentes de IA	4
2.3 Recuperação Aumentada por Geração (RAG)	5
2.4 Ferramentas	6
2.5 Aplicações de Agentes de IA no E-commerce	8
2.6 Lacunas na Literatura e Oportunidade de Pesquisa	8
2.7 N8N como orquestrador no ecossistema de agentes	8
3. Business Process Management (Gestão de Processos de Negócio):	9
3.1 Contexto:	9
3.2 Benefícios do BPM:	10
3.3 Importância no contexto empresarial:	11
3.4 Ciclo de vida BPM	11
4. Estudo de Caso:	14
4.1 Contexto organizacional e escopo	14
4.2 Descrição dos Processos (AS-IS)	15
4.3 Modelos BPMN (AS-IS)	20
4.4 Gargalos e diagnóstico	24
4.5 Objetivos da intervenção e perguntas de pesquisa	25
4.6 Redesenho do processo (TO-BE):	26
4.7 Modelos BPMN (TO-BE)	30
4.8 Plano de medição (KPIs, metas e hipótese de impacto)	32
5. Arquitetura da Solução, Dados e Implementação:	35
5.1 n8n como orquestrador do redesenho	35
5.2 Arquitetura lógica e fluxos ponta a ponta	37
5.3 Dados e Transformações	38
5.4 Ferramentas (tools)	38
5.5 Implementação e Governança do Agente de IA	39
5.6 Limitações e próximos passos	42
6. Avaliação, Resultados e Discussão:	43
6.1 Objetivo e desenho da avaliação	43
6.2 Coleta e análise	44
6.3 Limitações do experimento	44
6.4 Resultados e Discussão	45
6.5 Principais resultados	45
6.6 Interpretação dos resultados	49
7. Conclusão:	50
7.1 Trabalhos futuros	51

8. Referências Bibliográficas:	52
Glossário:	54
Apêndice A: Arquitetura Lógica da Solução:	56
Apêndice B: Resultado do Survey no Google Forms:	65

LISTA DE FIGURAS:

Figura 1 - O Ciclo de Vida BPM.....	13
Figura 2 - Arquitetura tecnológica atual da Brofitwear.....	14
Figura 3 - Processo de Compra.....	20
Figura 4 - Faturamento e Envio.....	21
Figura 5 - Logística Reversa.....	22
Figura 6 - Pós logística reversa.....	23
Figura 7 - Logística Reversa (TO-BE):.....	31
Figura 8 - Processo de Compra (TO-BE):.....	32
Figura 9 - Arquitetura da solução Resumida.....	36
Figura 10 - Estrutura de orquestração do agente no n8n:.....	38
Figura 11 - Pergunta do formulário: O agente foi fácil de usar?.....	50

LISTA DE TABELAS:

Tabela 1 - Problemas identificados pelos stakeholders por fluxo.....	24
Tabela 2 - Heurísticas de Redesenho e seus Objetivos.....	30
Tabela 3 - KPIs do piloto fechado (definição, meta, fonte, observações).....	34
Tabela 4 - Performance geral do agente.....	47
Tabela 5 - Uso de cada ferramenta.....	48
Tabela 6 - Execução geral do bot.....	48
Tabela 7 - Performance geral das ferramentas.....	49
Tabela 8 - Detalhamento das métricas de cada ferramenta.....	49

1. Introdução

O comércio, ao longo da história, se transformou de práticas de troca para sistemas complexos de circulação de mercadorias, informação e serviços. Atualmente, o comércio eletrônico se consolidou em um arranjo técnico em que operações, marketing, atendimento e logística são orquestrados digitalmente e em tempo real. Nesse ambiente, empresas de porte médio, com times enxutos e cadeias de integração ERP/CRM/OMS/WMS, enfrentam o desafio de alinhar eficiência operacional, qualidade de serviço e experiência do cliente em múltiplos pontos de contato ao longo da jornada do consumidor. A gestão por processos de negócio fornece o enquadramento conceitual para entender como essas organizações transformam conjuntos de atividades interdependentes em valor, deslocando o foco de funções isoladas para fluxos de trabalho do início ao fim. Essa perspectiva, difundida por abordagens como BPM e por notações como BPMN, se mostra muito pertinente a organizações intensivas em informação e com forte adoção a sistemas integrados. (ABPMP, 2013)

A evolução tecnológica recente reconfigura esse quadro ao introduzir agentes de Inteligência Artificial (IA) capazes de interagir em linguagem natural, acessar bases de conhecimento, acionar ferramentas e registrar decisões com rastreabilidade. Diferentemente de chatbots determinísticos, esses agentes combinam *Retrieval Augmented Generation* (RAG), chamadas a sistemas integrados (por exemplo, CRM, OMS e WMS) e mecanismos de auditoria, sempre operando sob instruções, políticas e *guardrails*. Em operações de *e-commerce*, eles podem atuar em frentes de atendimento *omnichannel*, autosserviço (WISMO, trocas e devoluções) e recomendações. Ao incorporar humano-no-circuito em processos sensíveis, empresas inovadoras que adotam a IA em seus processos buscam ganhos mensuráveis em conversão, ticket médio e NPS, preservando governança e segurança.

Nesse contexto se insere o objeto deste estudo: Uma empresa brasileira de médio porte no segmento de moda fitness, com operação predominante no varejo e presença secundária no e-commerce. Seus pontos críticos atuais se manifestam em três pontos: (i) personalização limitada da jornada do consumidor; (ii) autosserviço insuficiente; (iii) retrabalho entre canais. Se propõe investigar um redesenho do processo de atendimento e a adoção de um Agente de IA como sua solução, ancorada em BPM e alinhada aos princípios da Indústria 5.0: centrada no ser humano, sustentável, resiliente, personalizável e com cooperação homem-máquina.

Assim, compreendendo o agente como novo “recurso” integrado ao desenho organizacional e aos fluxos TO-BE.

A escolha por BPM como base de referência decorre de sua ênfase na melhoria contínua, no monitoramento por indicadores e na documentação sistemática (AS-IS/TO-BE) que viabiliza a transparência e aprendizado organizacional. Na literatura aplicada à gestão por processos, observamos sinergias entre mapeamento do processo, definição de metas e construção de painéis de indicadores para monitorar criticidade e desempenho, o que reforça a pertinência do enquadramento para o estudo de caso. (*ABPMP, 2013*)

Do ponto de vista metodológico, o trabalho se estrutura como estudo de caso único com evidências mistas. A estratégia compreende: (i) mapeamento AS-IS dos processos críticos relacionados às frentes de atuação do agente e desenho TO-BE com pontos de controle; (ii) implantação controlada do agente em fluxos selecionados; e (iii) avaliação do impacto do redesenho por meio de análise de dados operacionais e um experimento com um pequeno grupo de candidatos utilizando a nova ferramenta. Esse desenho segue recomendações de pesquisa empírica na Engenharia de Produção, valorizando validade interna/externa, critérios de decisão e utilização de métodos quantitativos e qualitativos de forma complementar.

No meio acadêmico, a introdução deste trabalho cumpre o papel de situar o fenômeno, explicitar sua relevância, indicar a questão de pesquisa e enunciar objetivos, métodos e estrutura. Assim, se pergunta: em que medida a adoção de um Agente de IA, desenhado e governado sob lógica BPM, melhora resultados de operação e experiência do cliente no e-commerce da empresa, à luz dos princípios da Indústria 5.0? A resposta a essa questão será buscada por meio dos objetivos (geral e específicos) estruturados ao problema, com delimitações, métricas e hipóteses quando possível.

O objetivo geral é analisar em que medida a introdução desse agente, desenhado e governado sob lógica de gestão de processos, melhora resultados operacionais e de experiência na empresa. Como objetivos específicos, se propõe: (i) mapear processos críticos AS-IS e projetar fluxos TO-BE; (ii) implementar o agente em sua fase piloto; (iii) avaliar seus impactos por meio de experimentos e análises de métricas internas.

Metodologicamente, se adota um estudo de caso único com evidências mistas, combinando análise documental, mapeamento BPMN, experimento e entrevistas com

stakeholders. Assim, se prevê as fases de diagnóstico, desenho, implementação controlada e avaliação, com foco em validação interna, rastreabilidade e documentação dos sistemas.

A organização deste trabalho é a seguinte. O Capítulo 2 apresenta o referencial teórico sobre agentes de IA, arquitetura de agentes, Recuperação Aumentada por Geração (RAG), uso de ferramentas (tools) e aplicações em e-commerce, incluindo o papel do n8n como orquestrador no ecossistema de agentes. O Capítulo 3 aprofunda a Gestão de Processos de Negócio (BPM), discutindo sua evolução histórica, benefícios, importância nas organizações e ciclo de vida, que servem de base para o redesenho proposto. O Capítulo 4 descreve o estudo de caso da Brofitwear: caracteriza o contexto organizacional e o escopo, apresenta os processos AS-IS em BPMN, identifica gargalos, explicita os objetivos da intervenção e as perguntas de pesquisa, e propõe o redesenho TO-BE com seu plano de medição.

O Capítulo 5 detalha a arquitetura da solução, dados e implementação, mostrando como o n8n, o agente de IA, as integrações VTEX/Correios e o repositório de conhecimento se articulam para materializar o TO-BE em um arranjo governado por processos. O Capítulo 6 apresenta o desenho da avaliação, o procedimento de coleta e análise dos dados, as limitações do experimento e os principais resultados e discussões. Por fim, o Capítulo 7 sintetiza as conclusões do estudo, discute implicações para a gestão e aponta trabalhos futuros, destacando como a combinação entre BPM, agentes de IA e integrações corporativas pode apoiar a evolução da operação da Brofitwear e de organizações semelhantes.

2. Agentes de IA, RAG, Ferramentas e suas Aplicações

2.1 Inteligência Artificial e Agentes Inteligentes

A Inteligência Artificial (IA) pode ser compreendida como o campo dedicado ao desenvolvimento de tecnologias capazes de realizar tarefas que, se desempenhadas por humanos, demandariam inteligência. Historicamente, sua trajetória percorre três grandes marcos: (i) sistemas baseados em regras e especialistas, centrados em conhecimento explícito e inferência simbólica; (ii) a ascensão do *machine learning* estatístico, com destaque para redes neurais e *deep learning*, sustentados por dados em escala; e (iii) a etapa atual de modelos generativos de larga escala (LLM), capazes de produzir linguagem natural, código, imagens e vídeos a partir de instruções por meio de texto. (BROWN et al., 2020)

Nesse contexto emergem os agentes inteligentes, aqui entendidos como sistemas com autonomia operacional (capacidade de decidir *o que* e *quando* fazer dentro de políticas definidas), proatividade (iniciar ações orientadas a objetivos), aprendizado (ajuste do comportamento a partir de dados/feedback) e interação (com pessoas, dados e serviços). Os agentes se diferenciam de *chatbots* convencionais por integrarem percepção, raciocínio e ação junto com processos organizacionais: não apenas respondem, mas consultam fontes, executam ferramentas, registram eventos e tomam decisões. (OPENAI et al., 2023)

A diferença operacional entre um *chatbot* simples e um agente de IA consiste, portanto, na profundidade de integração (sistemas transacionais, APIs, bases de conhecimento), na capacidade de tomada de decisão (políticas, *guardrails*, critérios) e no suporte a processos de ponta a ponta (*workflows*). Em ambientes empresariais, esses agentes podem apoiar operações, atendimento e análise de dados, ampliando eficiência, qualidade e experiência do cliente, desde que sejam inseridos em processos definidos, estáveis e mensuráveis.

2.2 Arquitetura de Agentes de IA

A arquitetura típica de um agente se compõe de três camadas:

- A “percepção” agrega entradas (texto, eventos, documentos, métricas), realizando normalização e enriquecimento do contexto.

- O “raciocínio” combina modelos generativos e cálculos determinísticos para interpretar intenções, recuperar conhecimento, aplicar políticas e planejar ações.
- A ação se materializa em respostas, atualizações de registros, chamadas a APIs e criação de evidências (logs, *traces*).

Para o pleno funcionamento e integração dessa arquitetura em processos, é importante que em momentos críticos haja um humano para supervisionar, impor ou verificar as ações do agente: o *human-in-the-loop* é um princípio estrutural na implementação de agentes de IA e dialoga com a Indústria 5.0 ao favorecer cooperação homem-máquina e tomada de decisão responsável. Casos difíceis, sensíveis ou de alto impacto são enviados para humanos, assim preservando transparência e auditabilidade. (BAI et al., 2022)

Nesse sentido, a governança e os *guardrails* incluem *logging* detalhado, auditoria, *rate limits*, anonimização quando cabível, políticas de retenção, monitoramento de vieses e medidas de segurança (autenticação, autorização, escopos mínimos das chaves). Assim, a integração com sistemas corporativos (ERP, CRM, OMS, WMS) ocorre por contratos de API, assegurando consistência e rastreabilidade de ponta a ponta. (BAI et al., 2025)

2.3 Recuperação Aumentada por Geração (RAG)

A RAG é uma técnica em IA que separa o conhecimento paramétrico do modelo (o que o LLM “sabe” por treinamento) do conhecimento não-paramétrico (documentos externos atualizados), obrigando o agente a consultar fontes antes e, quando necessário, durante a geração de texto. Com isso, se reduz a dependência de memória estática e suas alucinações diminuem, ao mesmo tempo em que se preserva rastreabilidade por citações da origem consultada na base de dados. (LEWIS et al., 2020)

O pipeline típico compreende três fases:

1. Ingestão e indexação: Conteúdos são coletados, normalizados (limpeza de HTML, tabelas), segmentados (*chunking*) e representados via *embeddings* densos. Se armazenam em um repositório vetorial (com metadados como versão, categoria e idioma).
2. Recuperação: A consulta do usuário pode ser reformulada (expansão dos termos e sinônimos), filtrada por metadados e enviada ao índice para retorno de informações na

base vetorial. Quando a precisão é crítica, se aplica a técnica de *re-ranking* para ordenar os trechos mais prováveis de conter a resposta.

3. Geração condicionada: O LLM recebe somente o contexto recuperado (com limites da quantidade de termos), e elabora a resposta citando as fontes.

Do ponto de vista técnico, a RAG exige decisões sobre granularidade de *chunk*, dimensão de *embedding*, quantidade de informação recuperada, latência e custo, além de mecanismos de atualização (incremental vs. *batch*) e versionamento como *canary releases* para evitar respostas com base em documentos antigos. (GOOGLE, 2025)

A avaliação de um sistema que utiliza RAG para a geração de respostas deve incluir métricas como Precisão@k/Recall@k, factualidade/*groundedness* e taxa de abstinência de resposta, além de amostragens humanas periodicamente para verificar utilidade e clareza (CHEN *et al.*, 2024). No *e-commerce*, a RAG habilita respostas contextualizadas sobre políticas de trocas/devoluções, catálogo e fichas de produto, e pode interpretar diferentes status logísticos. Quando combinada a integrações de API (OMS, transportador, preços/estoque), o agente passa a “falar com dados oficiais”: consulta suas instruções para “entender ” o contexto e “interrogar” as APIs com valores dinâmicos, conciliando utilidade e conformidade. (ZHAI, 2025)

2.4 Ferramentas

No contexto de agentes, “*tools*” são funcionalidades externas estruturadas em funções com contrato (entrada/saída bem definidos) que o agente pode invocar para consultar dados, executar ações e produzir efeitos no mundo real. Em termos práticos, a *tool* é a “ponte” entre o modelo de linguagem e sistemas, sensores ou serviços: o LLM decide se e como chamá-la, fornece parâmetros (geralmente em *JSON*), aguarda a resposta e então integra o resultado ao raciocínio e à próxima ação. Esse mecanismo, também chamado de *function/tool calling*, transforma o agente de um gerador de texto em um orquestrador de processos.

Exemplos típicos de tools incluem: busca em índice vetorial (RAG) ou mecanismo web; consulta a catálogos e OMS (estoque, pedido, status); rastreamento logístico (SRO); envio de e-mails ou mensagens calculadora e utilidades (*parse*/normalização de dados). Cada *tool* deve declarar:

- (i) nome e propósito;

- (ii) esquema de entrada (tipos, obrigatoriedade, faixas);
- (iii) esquema de saída (campos, códigos de erro);
- (iv) pré-condições (ex.: “exige ID”);
- (v) efeitos colaterais (ex.: “abre solicitação”).

O ciclo de uso de *tools* costuma seguir três etapas:

1. Seleção (o agente decide qual tool atende à intenção, possivelmente guiado por regras e políticas de prompt);
2. Preparo de parâmetros (normalização de texto, URL-*encode*, validações);
3. Execução e interpretação (chamada, tratamento de erros, e sumarização do retorno).

Em nosso estudo, o catálogo de *tools* reflete as necessidades do *e-commerce*:

- Procurar Produto (VTEX Search): Recebe consulta normalizada, retorna candidatos de peças adequadas “ao vivo”; usado junto de filtro por categoria e *re-ranking* (Cohere) para priorizar relevância.
- Base de Dados (Supabase, RAG): Recupera trechos de fichas de produto e políticas/FAQ, servindo de fonte primária antes do live search na VTEX.
- Procurar Pedido (VTEX OMS): Fonte canônica de status e promessas de entrega.
- Solicitação Correios (logística reversa): Emissão autorização de postagem reversa nos sistemas de correios.
- Mandar E-mail de Suporte: Escalonamento para o SAC no caso de exceções.

Essas *tools* são orquestradas por um roteador: Regras determinísticas asseguram conformidade, enquanto a camada semântica classifica a intenção do usuário (recomendação, FAQ/política, *Where Is My Order?* Logística Reversa). Por fim, *guardrails* no *prompt* restringem comportamentos (não inventar links/dados, citar fontes internas) e observabilidade (*logs*) registra decisões, parâmetros e resultados, permitindo medir aderência às regras, erros de validação, tempos de ciclo e taxa de fallback para humano, requisitos essenciais ao uso responsável de agentes em processos de negócio.

2.5 Aplicações de Agentes de IA no E-commerce

No atendimento *omnichannel*, agentes possibilitam personalização em tempo real, FAQ dinâmico, recuperação de carrinho e resoluções *self-service* para *WISMO* e pós-venda.

Em logística e operações, executam consultas de status, iniciam fluxos de devolução, verificam disponibilidade. Em marketing e vendas, suportam segmentação, *cross-sell/up-sell* e recomendação contextualizada. A efetividade do agente em cada um desses processos depende do quão configurado foi o *bot* ao processo, do desenho de *guardrails* e da disciplina de medição. (MAO; YANG; FU, 2025)

2.6 Lacunas na Literatura e Oportunidade de Pesquisa

Atualmente na literatura sobre o uso de inteligência artificial na indústria, se observa grande ênfase de casos e relatos de grandes organizações e plataformas globais, com menor cobertura de empresas médias, com perfil predominante no varejo digital. Além disso, a integração entre BPM e agentes de IA em *e-commerce* permanece pouco explorada, principalmente em relação a governança, privacidade, *compliance* e integrações em tempo real. Essas limitações abrem espaço para estudos que avaliem, com método, o valor gerado por agentes em operações digitais, conectando desenho de processos (AS-IS/TO-BE), arquitetura técnica, implantação controlada e análise de impacto.

2.7 N8N como orquestrador no ecossistema de agentes

O n8n é uma plataforma de automação *low-code* orientada a fluxos que permite orquestrar integrações, transformar dados e encadear ações entre diferentes sistemas por meio de *nodes* (conectores) e *triggers* (*webhooks*, eventos). No contexto de agentes de IA, cumpre o papel de camada de execução operacional: o LLM decide “o que” fazer (intenção e *tool*), enquanto o n8n executa “como” fazer (autenticação, chamadas a APIs, validações, *logging* e políticas de erro), preservando governança e rastreabilidade. (N8N,2025)

Do ponto de vista conceitual, o n8n contribui para a engenharia de processos ao materializar, em fluxos explícitos, regras de negócio, pré-condições e efeitos de cada ação do agente. Isso o aproxima das práticas de BPM: cada *workflow* torna visíveis as entradas/saídas, exceções e pontos de controle, favorecendo o ciclo “modelar, implementar, monitorar, melhorar”. Esse sistema, também dialoga com a Indústria 5.0, ao permitir cooperação homem-máquina: quando o fluxo atinge um caso sensível, o n8n aciona o humano-no-circuito (e-mail para o Suporte) com contexto suficiente para continuidade.

3. Business Process Management (Gestão de Processos de Negócio)

3.1 Contexto

Processos são a espinha dorsal de qualquer esforço humano organizado: sequências de atividades que transformam insumos em resultados, replicáveis e passíveis de melhora. Historicamente, seu aperfeiçoamento decorreu tanto da experiência prática, o “aprender fazendo”, quanto da adoção de métodos científicos e tecnológicos. Em um mercado globalizado, marcado por competição intensa, ciclos curtos e exigências de conformidade, a otimização de processos deixou de ser vantagem opcional para se tornar condição de sobrevivência de uma organização.

O que chamamos hoje de Business Process Management (BPM) não surgiu de forma inusitada, é síntese de diversas contribuições acumuladas. Adam Smith, ao discutir a divisão do trabalho, evidenciou os ganhos decorrentes da especialização e do arranjo das tarefas ao longo de um fluxo produtivo. Frederick Taylor consolidou a mensuração, a padronização e o controle como fundamentos de eficiência industrial, inaugurando uma abordagem sistemática para decompor e redesenhar o trabalho. No pós-guerra, W. Edwards Deming apresentou ciclos de melhoria contínua e gestão por qualidade, conectando dados, variação e aprendizado organizacional. Taiichi Ohno, por sua vez, estabeleceu bases da produção enxuta, com foco na eliminação de desperdícios e fluxo contínuo, influenciando a gestão por processos além do chão de fábrica. (*A TRAJETÓRIA DO BPM, s.d.*)

Nos anos 1990, Michael Hammer popularizou a Reengenharia de Processos, defendendo revisões radicais de ponta a ponta, orientadas ao cliente e suportadas por tecnologia da informação. Já no início dos anos 2000, Peter Fingar e Howard Smith sistematizaram o BPM como disciplina integrada e orientada a desempenho. O livro *Business Process Management - The Third Wave* (2002) é frequentemente citado como marco da terceira onda, ao articular princípios, métodos e uma visão tecnológica que extrapola o mapeamento para discutir sobre execução, monitoramento e transformação contínua. (*SMITH; FINGAR, 2007*)

A institucionalização do BPM acompanha esse movimento. A ABPMP (Association of Business Process Management Professionals) foi fundada em 2003 e consolidada globalmente em 2010, difundindo o BPM CBOK®, guia para o corpo comum de conhecimento em BPM.

Atualmente o ramo se estrutura em três pilares: o BPM CBOK®, que organiza áreas como modelagem, análise, desempenho, transformação e governança; a BPMN, notação padrão que fornece linguagem comum e precisão analítica aos modelos; e os BPMS, plataformas que executam, monitoram e orquestram processos em tempo real. Em conjunto, esses elementos articulam teoria, prática e ferramentas para transformar a visão de processos em uma organização mensurável e sustentável. Em suma, o BPM integra aprendizados históricos - Qualidade, melhoria contínua, reengenharia, Lean e Six Sigma - Em uma disciplina moderna, capaz de alinhar estratégia, trabalho e tecnologia na busca de desempenho superior.

3.2 Benefícios do BPM

BPM é, acima de tudo, gestão orientada a resultados: desenhar, operar, medir e transformar processos para entregar valor com menor custo e maior previsibilidade. Nesse sentido, podemos defini-lo como uma disciplina que gera processos melhores ao menor custo total, articulando pessoas, insumos, regras e softwares, e criando condições estruturais para adaptação contínua e liderança em mercados inovadores.

Quando processos são descritos e governados, o seu crescimento deixa de ser um salto arriscado e se torna uma consequência: Operar em múltiplos locais ou canais e incorporar inovações passa a integrar a rotina. A modelagem, conforme o BPM CBOK, fornece habilidades, técnicas, tipos de modelos e padrões (como BPMN) para compreender e comunicar o funcionamento ponta a ponta, reduzindo ambiguidades e promovendo decisões baseadas em evidências. (*ABPMP, 2013*)

Os benefícios se distribuem por diferentes stakeholders. Para a organização, se destacam: responsabilidades claras; métricas e resposta ágil; controle de custos e qualidade; conformidade e governança com menor custo; visibilidade do fluxo de valor e prontidão para mudanças; acesso à informação estruturada; avaliação de custos; consistência operacional e gestão do conhecimento. Para o cliente, processos transformados significam atendimento mais rápido e confiável, cumprimento de promessas e foco no que agrega valor. Para a gerência, o BPM viabiliza otimização ponta a ponta, melhor planejamento de capacidade, benchmarking e gestão de incidentes. Para os atores de processo, há clareza de papéis, visão do todo, definição de requisitos do trabalho, ferramentas adequadas e reconhecimento pelo desempenho.

Do ponto de vista do sistema, medir e monitorar o desempenho de processos, não apenas de áreas, é uma condição essencial para melhoria contínua e controle efetivo. Procedimentos claros e monitoramento reduzem custos de compliance e riscos. O mapeamento viabiliza custeio por atividades, redução de desperdícios e precificação mais precisa. Ao aprofundar o conhecimento do negócio, se cria uma base para padronização, capacidade e sustentabilidade. Dessa forma, uma transformação bem sucedida exige alinhamento entre pessoas, tecnologia e processos, com mudanças culturais, técnicas e metodológicas, apoiadas por definição de nível de excelência e priorização de melhorias. O resultado é maior agilidade, produtividade, desempenho operacional e qualidade percebida pelo cliente.

3.3 Importância no contexto empresarial

Num contexto empresarial, o BPM dá uma visão detalhada dos fluxos de ponta a ponta, permitindo analisar causas, propor mudanças e institucionalizar a melhoria contínua. Ao disciplinar o ciclo “medir-aprender-intervir”, viabiliza ganhos de produtividade, qualidade e economia tanto de pessoal, materiais e financeiros.

Ao padronizar atividades, o BPM reduz ambiguidade e variabilidade, tornando mensuráveis os resultados operacionais. Essa padronização sustenta indicadores confiáveis que, alinhados a metas estratégicas, facilitam comparações internas (*benchmarking*) entre unidades, canais ou turnos. Com dados de processo rastreáveis, se tornam possíveis análises preditivas (capacidade, demanda, risco), deslocando a organização de uma postura reativa para estratégias mais agressivas de mercado.

Os resultados de negócio decorrem dessa disciplina: redução de custos, aumento de receitas, maior eficiência operacional e abertura de oportunidades de crescimento. Em termos gerenciais, o BPM cobre o ciclo completo: Da estratégia em processos e metas; modelagem e planejamento; execução e orquestração; monitoramento de performance; relatórios e análise para decisão. (ABPMP, 2013)

3.4 Ciclo de vida BPM

Na literatura, o ciclo de vida do BPM organiza a gestão por processos em etapas que vão do planejamento à melhoria contínua. Adotamos aqui seis fases: identificar, descobrir,

analisar, desenhar, implementar e monitorar. Abaixo vemos o detalhamento de cada uma dessas etapas e a Figura 1, que ilustra exatamente esse processo cíclico. (DUMAS *et al.*, 2018)

Identificar Processos: Se define a arquitetura de processos (cadeia de valor, macroprocessos e relacionamentos), suas fronteiras, stakeholders e donos de processo (process owners), além de critérios de priorização (risco, impacto, aderência estratégica). O resultado é um mapa de “o que importa” para a organização e onde concentrar esforços de descoberta e melhoria.

Descobrir Processos (AS-IS): Se levantam como os processos realmente funcionam hoje, por meio de documentos, entrevistas, observação, dados operacionais e modelagem (p.ex., BPMN nos níveis adequados). Se capturam eventos, regras, dados de entrada/saída, variações e exceções; quando possível, se estabelece uma baseline de desempenho (tempo, custo, qualidade).

Analisar: Com o processo AS-IS explicitado, se coletam dados (tempo, custo, qualidade, volume), se mapeiam variações e gargalos e se investigam causas-raiz. A modelagem torna o fluxo observável e comunicável entre áreas; os textos didáticos de BPM destacam essa etapa como base para qualquer intervenção subsequente, pois transforma percepções em evidências. (ABPMP, 2013)

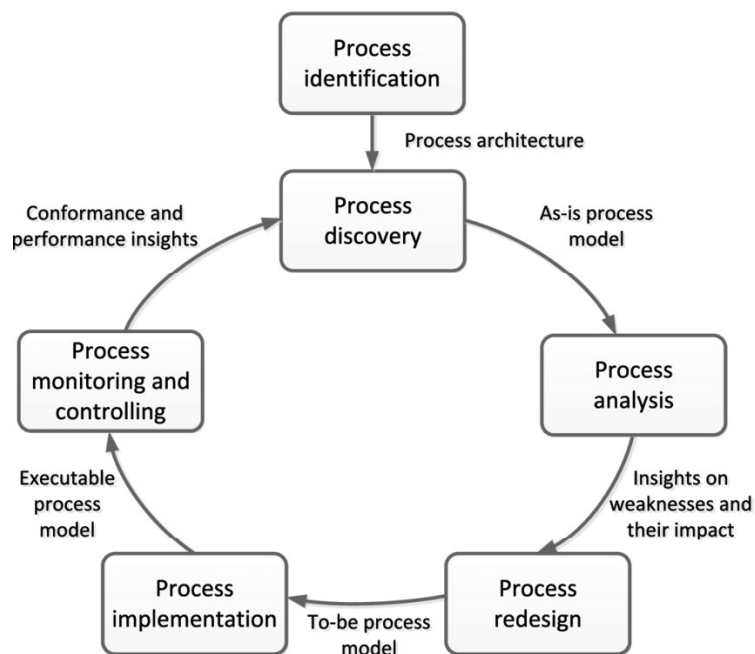
Desenhar: Se constrói o TO-BE aplicando diretrizes de redesenho (eliminação de atividades sem valor, paralelização, balanceamento de carga, *straight-through processing* quando pertinente), avaliando *trade-offs* em custo, tempo, qualidade e flexibilidade. A notação BPMN 2.0.2 é recomendada como linguagem comum para especificar o novo fluxo com precisão e portabilidade entre ferramentas. (REIJERS; LIMANMANSAR, 2005)

Implementar: A transição do desenho para a execução combina ajustes organizacionais (papéis, treinamento, regras) e tecnológicos (integrações, BPMS, orquestração). O ensino de BPM enfatiza a automação onde houver estabilidade/baixo desvio padrão, mantendo *guardrails* para exceções e garantindo rastreabilidade das mudanças.

Monitorar: Em operação, se mede o desempenho do processo (não apenas de áreas) com indicadores de eficácia e eficiência, *dashboards* e registros de eventos. O monitoramento

fecha o ciclo “medir-aprender-intervir”, permitindo detectar desvios, comprovar ganhos e sustentar governança de ponta a ponta ao longo do tempo.

Figura 1 - O Ciclo de Vida BPM



Fonte: (DUMAS et al., 2018)

Em síntese, as seis fases formam um ciclo gerencial completo: Da intenção estratégica à evidência operacional, que estrutura a trajetória do mapeamento até a melhoria sustentada.

4. Estudo de Caso

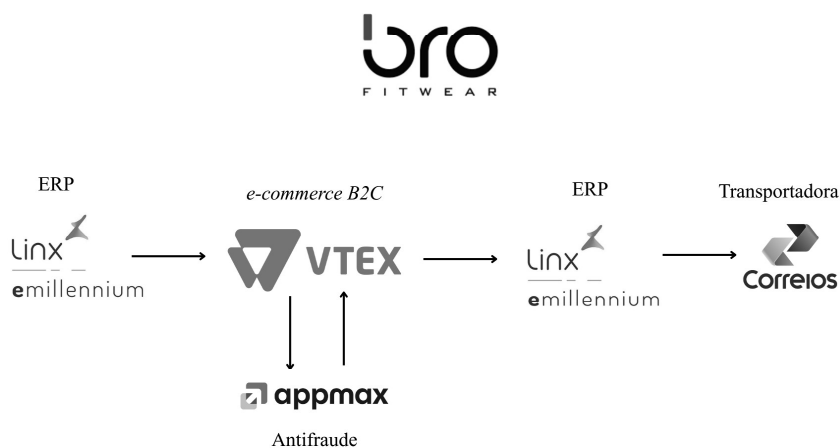
4.1 Contexto organizacional e escopo

A Brofitwear atua no varejo de moda fitness, com portfólio centrado no público feminino com linhas complementares masculina e de acessórios em expansão. O quadro atual soma aproximadamente 80 colaboradores. O canal digital se encontra em crescimento, mas ainda representa parcela minoritária do faturamento anual, uma vez que mais de 50% das receitas advêm de revendedores no Brasil e no exterior. A operação comercial é multicanal: site próprio, rede de lojas multimarcas (revendedores) e presença física em RJ, SP, Porto Alegre e Bahia. A marca atende clientes no Brasil e, por meio de revendedores, em mercados dos Estados Unidos, Europa e África. O público-alvo principal se concentra em mulheres de 25-35 anos, com alta afinidade à saúde e prática de atividades físicas.

Do ponto de vista tecnológico, o site e as vendas são orquestrados pela VTEX; o sistema ERP é o e-Millennium (Linx); e o antifraude é realizado via Appmax. Em logística, os Correios são o operador predominante, com coleta diária na fábrica sob contrato. A partir desse *stack* tecnológico, a solução para alguns processos pouco eficientes encontrados será discutida nos capítulos seguintes, integrando o *stack* utilizando IA, a ferramenta n8n (para orquestração), Supabase (vetores, memória e logs), VTEX API e Correios API como motores principais para o redesenho de processos.

A Figura 2 abaixo, explicita o fluxo de informação e sistemas internos que operam para garantir uma operação online robusta na empresa

Figura 2 - Arquitetura tecnológica atual da Brofitwear



Fonte: Autoria Própria

O escopo deste estudo cobre a jornada completa do consumidor, do pré-venda/*discovery* ao pós-venda: *checkout*/pagamento, faturamento, picking/packing/expedição, entrega (*WISMO*), trocas/devoluções e atendimento/SAC. A opção por priorizar essa jornada, se apoia em quatro razões:

1. Patrocínio e prioridade executiva: houve demanda direta da direção de Marketing para avaliar o impacto operacional e comercial de intervenções com IA na experiência do cliente.
2. Alinhamento estratégico e alavancas de valor: o canal digital se encontra em crescimento, mas com uma baixa conversão de 0,50% e ticket médio de R\$390; e fortalecer o *e-commerce* é objetivo explícito diante da dependência atual de revendedores em lojas multi-marcas. A jornada do cliente concentra métricas de negócio sensíveis (conversão, AOV) e custos de serviço (*WISMO*, tempo de ciclo), permitindo capturar valor mensurável em curto prazo.
3. Viabilidade técnica: Há dados e integrações disponíveis (VTEX/OMS e Search; Correios/SRO; n8n; Supabase) que viabilizam a implementação de agentes de IA com RAG e *tools* transacionais, reduzindo incerteza técnica.
4. Horizonte e risco do projeto: o tempo reduzido de implantação e teste recomenda focar em um processo transversal, de alto volume e ciclo curto, com *feedback* rápido.

Nesse contexto, o objetivo geral da intervenção é apoiar o processo de compra e o atendimento para elevar a conversão, aumentar o ticket médio, reduzir contatos *WISMO* e encurtar o tempo de ciclo. Como restrição, se reconhece que a janela curta limita a robustez de algumas análises e por isso, o estudo adota métricas operacionais instrumentadas, experimentos e documentação sistemática para sustentar a validade dos resultados.

4.2 Descrição dos Processos (AS-IS)

Para compreender de forma aprofundada o funcionamento da jornada do consumidor e seus impactos na eficiência operacional, foi realizada uma análise detalhada dos processos internos que compõem o funil de compra e pós-venda. Essa etapa teve como objetivo mapear a cadeia de valor, delimitar o escopo e identificar os principais participantes, fluxos e integrações que sustentam a operação do *e-commerce*. A descrição dos processos AS-IS (situação atual) é fundamental para revelar gargalos ou redundâncias entre áreas e sistemas,

oferecendo uma base concreta para o diagnóstico de desempenho e o redesenho apresentado nos capítulos seguintes. A partir dessa análise, se tornou possível visualizar com clareza como as atividades se relacionam e quais oportunidades podem ser exploradas pela introdução do agente de IA como instrumento de integração e melhoria contínua.

4.2.1 Participantes, responsabilidades e sistemas (piscinas/raias)

No arranjo da operação, cinco participantes cumprem papéis complementares e se articulam por meio de mensagens e eventos. O Cliente é tratado como uma “caixa-preta” do ponto de vista do processo: inicia a compra no site, fornece os dados necessários ao *checkout*, recebe notificações transacionais e, no pós-venda, solicita devolução/troca e escolhe a tratativa (cupom ou reembolso).

O Website/*E-commerce* (VTEX/OMS) concentra a experiência de compra e a orquestração transacional: captura os dados de checkout, aciona o antifraude (Appmax), consolida a decisão de pagamento (aprovar/recusar), cria a ordem no OMS e publica o estado do pedido ao longo do ciclo. A partir do “pagamento aprovado”, o sistema dispara o fluxo interno de expedição e mantém a “fonte canônica” de *status*, consultada pelo atendimento e pelo agente de IA.

O setor de expedição transforma a ordem em produto expedido: executa *picking*, *packing*, emissão de NF-e, etiquetagem e expedição; além disso, efetiva a baixa financeira/reembolso quando aplicável. O *status* “pedido entregue” é registrado no ERP após a confirmação logística, fechando o ciclo operacional.

Os Correios (transportador) assumem a responsabilidade após a coleta diária: movimentam a remessa, atualizam o rastreamento (SRO) e, na logística reversa, emitem/cancelam autorizações de postagem conforme as regras de validade. Seus eventos alimentam o OMS/ERP e disparam comunicações ao Cliente. (CORREIOS, 2025)

Por fim, o Gateway/Antifraude (Appmax) opera como serviço externo especializado de risco, retornando aprovação ou negativa ao sistema interno da VTEX, que então continua (ou encerra) o fluxo. Em termos de comunicação, as integrações seguem pontos de controle claros: “pagamento aprovado” aciona faturamento/expedição; “postado/em trânsito/entregue” dispara notificações ao Cliente e atualização contábil; “autorização de reversa emitida/expirada” inicia prazos e instruções no pós-venda. Esse encadeamento, representado em BPMN no capítulo 4, estrutura responsabilidades e o caminho crítico da jornada.

4.2.2 Principais fluxos

(A) Processo de compra: O fluxo representado pela Figura 3 se inicia com o acesso do cliente ao site e a inclusão de itens no carrinho. A partir daí, o *website* verifica se o usuário está autenticado; caso não esteja, conduz ao cadastro/*login* e, em seguida, à escolha da modalidade de frete. Concluída essa etapa, o cliente informa os dados de pagamento e a transação segue para análise do serviço antifraude (Appmax), que retorna a decisão. Quando aprovada, se procede à reserva do estoque, à criação do pedido no OMS e ao envio da confirmação de compra ao cliente, se encerrando o fluxo de compra com a “entrega” do pedido ao processo subsequente de faturamento. Quando negada, o sistema notifica a não aprovação e a jornada se encerra sem geração de pedido. Se observa, no AS-IS, que a reserva de estoque ocorre apenas após a aprovação, uma escolha que reduz risco de ruptura.

(B) Processo de faturamento e envio: Como descrito na Figura 4, este fluxo é acionado pela mensagem “pagamento aprovado”. Internamente, a empresa executa a sequência operacional de *picking* (separação), *packing* (conferência e embalagem), emissão de NF e no ERP, impressão e aplicação da DANFE e da etiqueta de transporte, culminando na expedição. O cliente recebe a notificação de “pedido faturado” e, a partir da coleta diária, os Correios passam a registrar os eventos de trânsito no SRO até a confirmação de entrega. Com o fechamento logístico, o ERP registra “pedido entregue”, encerrando o ciclo. No AS-IS, se trata de um fluxo linear, sem gateways explicitamente mapeados para exceções como rejeição de NF, falha de coleta ou devoluções em trânsito, e há dependência da confirmação final dos Correios para baixa no ERP.

(C) Processo de logística reversa: Esse processo, como traduzido pelo BPMN na Figura 5, se inicia quando o consumidor solicita devolução/troca pelos canais disponíveis. A empresa, então, registra o pedido junto aos Correios por webservice e recebe uma autorização de postagem com validade de 10 dias. Os Correios emitem a autorização, cancelam-na automaticamente se o prazo expirar, recebem o objeto na agência e atualizam as informações de solicitação e de rastreamento. O consumidor, por sua vez, aguarda a autorização, realiza a postagem e acompanha a trajetória do objeto até o reingresso na empresa. Um ponto crítico no AS-IS é a verificação de retorno do número da etiqueta: quando presente, a empresa consegue rastrear o item; quando ausente, precisa acionar consultas adicionais (acompanhaPedido) até a regularização. Nesse processo existe risco de expiração da autorização, gerando cancelamentos e retrabalho.

(D) Processo pós-logística: Recebido o produto devolvido, a empresa realiza a conferência física para avaliar o estado do item. Representado na Figura 6, a decisão subsequente é binária: se o produto está em bom estado, o cliente é notificado do aceite e escolhe como proceder: cupom de valor equivalente no site (o qual é gerado e comunicado) ou reembolso direto (executado no financeiro/ERP); se o produto está danificado ou em desacordo com a política, o cliente é notificado da impossibilidade de troca/devolução e o item é devolvido via Correios, se dando o encerramento do caso. No AS-IS, o cliente surge predominantemente como receptor das decisões, sem checkpoints de transparência sobre prazos de análise, nem SLAs explícitos para reembolsos; tampouco há alternativas intermediárias (como reembolso parcial), o que limita a flexibilidade do atendimento em cenários ambíguos.

Em síntese, os quatro fluxos descritos formam a trilha funcional da jornada de compra e pós-compra. Embora sejam adequados para a operação atual, os modelos AS-IS apresentam oportunidades de enriquecimento, sobretudo na representação de exceções, nos pontos de decisão e na introdução de regras e métricas de desempenho, que serão objeto do redesenho (TO-BE) nas seções subsequentes.

4.2.3 Pontos de controle e integrações (onde medir)

Para sustentar o diagnóstico e orientar o redesenho, a operação estabelece pontos de controle no ERP que marcam transições ao longo do processo. Essas informações podem ser extraídas das integrações VTEX/OMS, Correios (SRO/Logística Reversa), ERP e n8n, e funcionam como “logs” do processo, a partir dos quais se calculam diferentes métricas relevantes a cada processo.

Na compra, o retorno do antifraude (“pagamento aprovado” ou “negado”) é o gatilho de medição. A partir desse evento, se monitora a taxa de aprovação, o abandono no pagamento e o tempo de autorização, indicadores diretamente associados à conversão e à experiência de *checkout*.

No faturamento e envio, dois eventos estruturam o acompanhamento: o registro de “pedido faturado” (publicado pela VTEX) e os eventos de “em trânsito/entregue” (publicados pelos Correios). Entre eles se calcula o tempo pedido até a postagem (horas) e, por meio da coleta diária, o percentual de coletas realizadas no tempo previsto. Com a confirmação

logística, se apuram OTIF (On Time In Full) e variações por região/CEP, insumos essenciais para ajustar promessas de entrega e alocar capacidade.

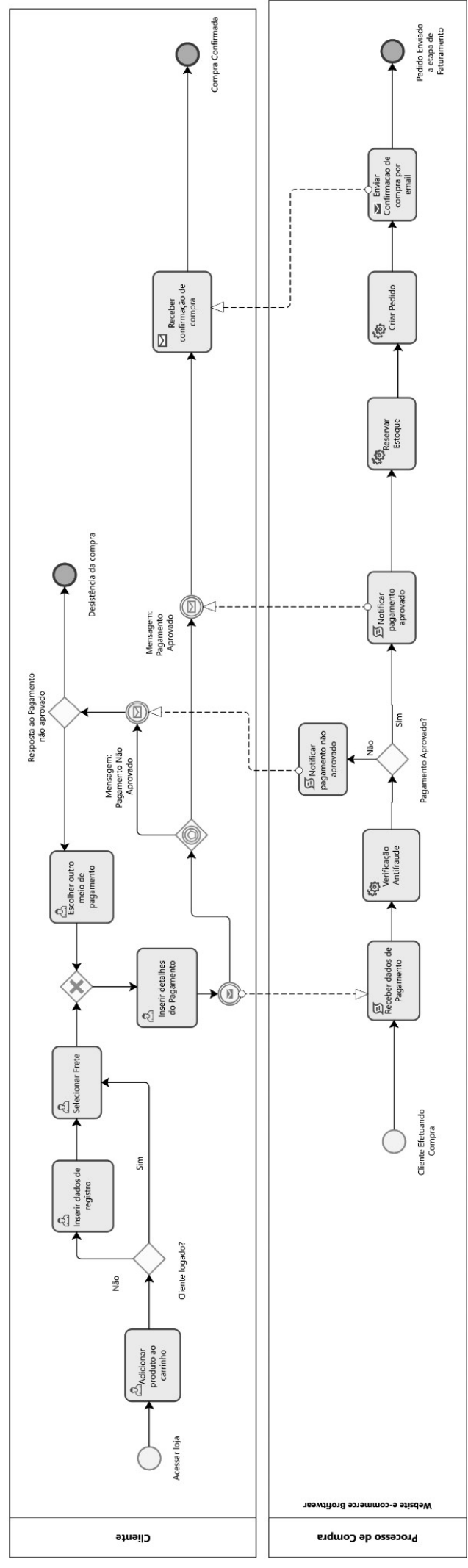
Na logística reversa, os eventos “autorização emitida/expirada” e “postagem confirmada” delineiam o funil de retorno. Desses marcos derivam a taxa de expiração (validade de 10 dias) e o lead time de retorno (solicitação até o recebimento), útil para dimensionar custos de reprocesso e políticas de crédito.

No pós-logística, a “conferência concluída” e a execução de “cupom/reembolso” encerram o caso e viabilizam a mensuração do prazo efetivo de reembolso/troca (baseline atual: 30 dias). A visibilidade desses tempos permite identificar gargalos na análise física, validações financeiras e comunicação com o cliente.

Todos esses eventos alimentam os sistemas da Linx e VTEX, consolidando métricas de processo, para suportar governança, auditoria e melhoria contínua.

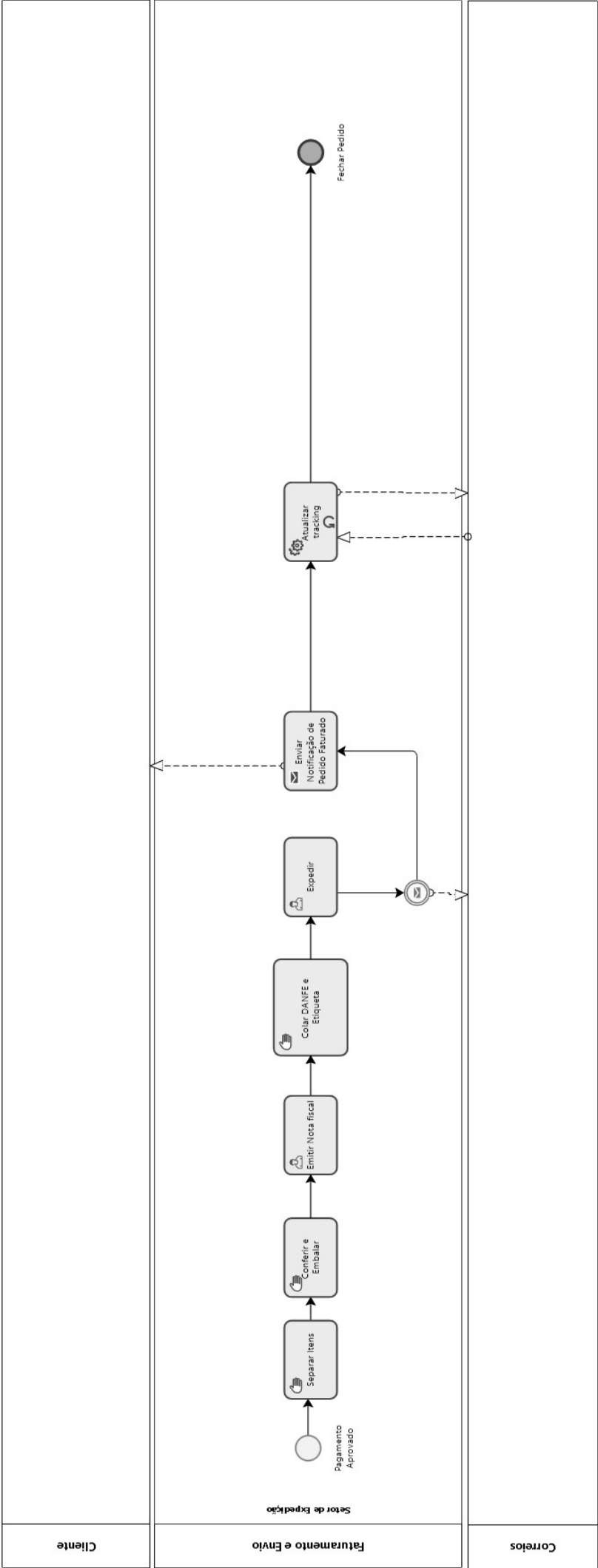
4.3 Modelos BPMN (AS-IS)

Figura 3 - Processo de Compra



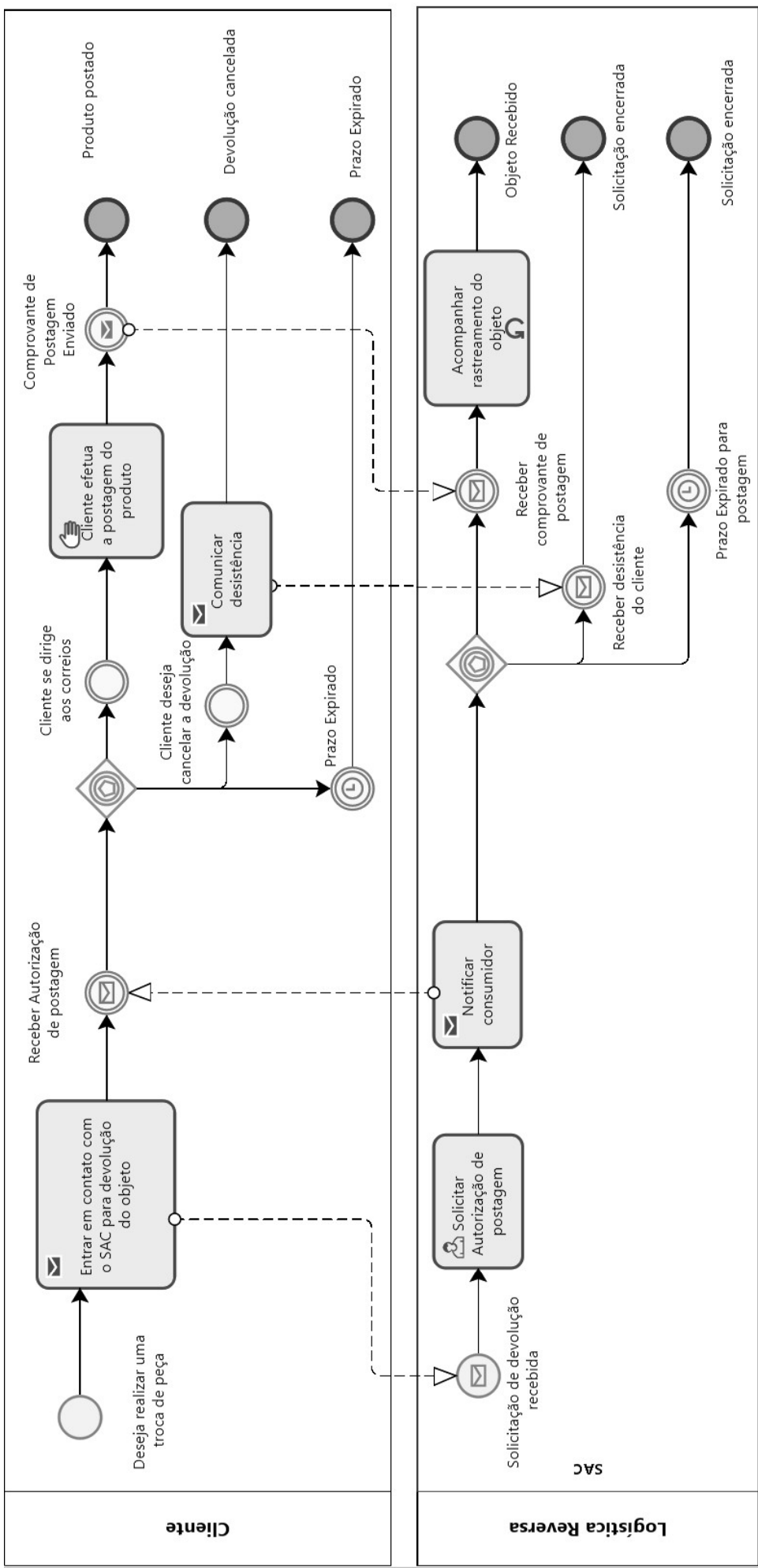
Fonte: Autoria Própria

Figura 4 - Faturamento e Envio



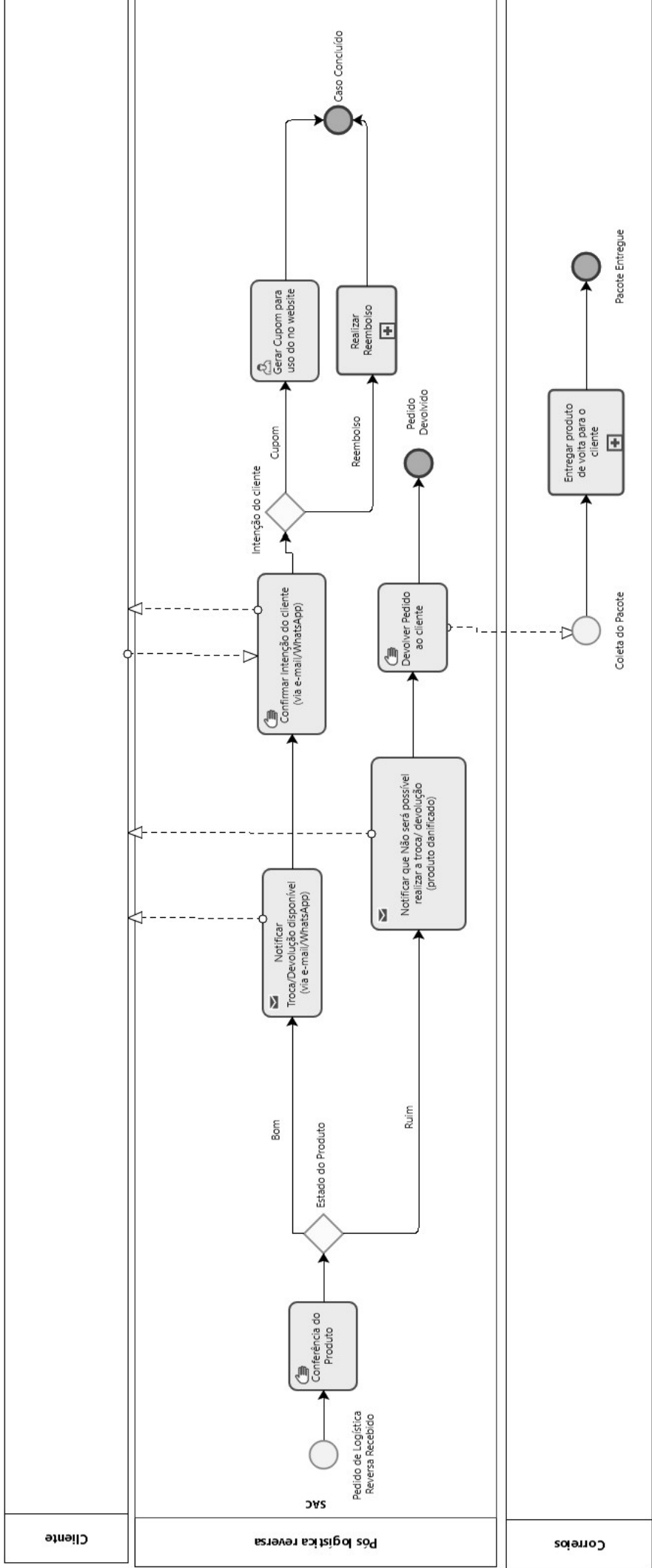
Fonte: Autoria Própria

Figura 5 - Logística Reversa



Fonte: Autoria Própria

Figura 6 - Pós logística reversa



Fonte: Autoria Própria

4.4 Gargalos e diagnóstico

A análise dos processos AS-IS combinou três fontes qualitativas principais:

- (i) entrevistas com gestores de Marketing, TI e SAC
- (ii) observação direta dos fluxos operacionais integrados à VTEX e ao ERP
- (iii) evidências relatadas durante o mapeamento em BPMN.

Para evitar confundir gargalos com defeitos, distinguimos:

- Gargalo: restrição estrutural ou etapa que limita o desempenho do fluxo.
- Problema: falha, erro, inconsistência ou variação indesejada em uma atividade específica.

A partir das entrevistas e do mapeamento AS-IS, consolidamos uma lista de problemas percebidos pelos stakeholders, organizada por fluxo da jornada do consumidor.

4.4.1 Lista de Problemas Relatados

Tabela 1 - Problemas identificados pelos stakeholders por processo

Processo	Problema relatado	Sintoma observado	Impacto percebido
Compra/Checkout	Dificuldade em encontrar peça ideal	Abandono de carrinho	Redução de conversão
Faturamento / Picking / Packing	Esperas na fila de faturamento	Atraso entre pagamento e expedição	Ampliação do lead time
Entrega / WISMO	Atrasos dos Correios	Cliente sem atualização	Crescimento de contatos WISMO
Logística Reversa	Comunicação com cliente lenta	Clientes insatisfeitos	Reclamações
Pós-logística	Processo manual de apuração das peças	Percepção de lentidão	Reclamações / follow-up
Transversais	Ausência de regras explícitas	Decisões ad hoc	Variabilidade, inconsistência

Fonte: entrevistas e observações operacionais (Autoria própria).

4.4.2 Causas-raiz prováveis

1. Processo:

- *Handoffs* sem SLA definido
- Ausência de gateways de exceção
- Falta de instruções operacionais padronizadas

2. Sistemas:

- Navegação no site complexa e extensa

3. Capacidade:

- Limitação física de estações de faturamento
- Coletas diárias que criam janelas rígidas para expedição

4. Políticas / Comunicação:

- Comunicação predominantemente reativa (WISMO e Reversa)

4.5 Objetivos da intervenção e perguntas de pesquisa

Após entrevistas com diferentes setores da empresa e principalmente a Direção Geral e de Marketing, a intervenção proposta combina redesenho de processos e um agente de IA para atingir quatro resultados-chave:

- (i) elevar a conversão
- (ii) aumentar o ticket médio
- (iii) reduzir os contatos WISMO
- (iv) encurtar o tempo de ciclo de compra.

Objetivos específicos por etapa da jornada do consumidor:

1. Pré-venda/Descoberta: aprimorar a relevância de recomendações e a orientação ao produto (via agente), com foco em engajamento qualificado que impacte conversão e ticket médio.
2. Checkout/Pagamento: reduzir fricções informacionais (dúvidas de política/entrega) que levam ao abandono.
3. Pós-venda/CS: diminuir a % de contatos WISMO e elevar FCR, por meio de autosserviço orientado por IA.

4. Devoluções/Trocas: aumentar o autoatendimento (abertura de logística reversa com o agente) e reduzir prazos operacionais associados.

Nesse sentido, o agente executará: (i) recomendações de produtos e assistência consultiva; (ii) respostas sobre políticas/FAQs e dados de produto; (iii) rastreamento de pedidos; (iv) abertura de postagem reversa nos Correios, com comunicação automatizada a clientes e ao time de suporte.

Perguntas de pesquisa:

- P1. Em que medida o redesenho + agente de IA aumenta a conversão e eleva o ticket médio no site?
- P2. O agente reduz a proporção de contatos WISMO e eleva o FCR no período observado?
- P3. O agente contribui para reduzir o tempo de resposta e de resolução no atendimento, melhorando a agilidade e a consistência das interações com o cliente?
- P4. Quais são os efeitos colaterais e *trade-offs* observados (latência do bot, indisponibilidade de serviços externos, cobertura de catálogo nas recomendações)?
- P5. Como a integração do agente com as fontes oficiais de dados (VTEX, Correios e Supabase) impacta a confiabilidade e precisão das respostas fornecidas ao cliente?

4.6 Redesenho do processo (TO-BE)

Partindo do diagnóstico AS-IS, os novos processos TO-BE ilustrados na Figura 7 e 8 adotam heurísticas de redesenho orientadas à tecnologia, focadas nas integrações de API da VTEX (OMS/Catálogo) e Correios (SRO e Logística Reversa), sendo orquestradas pelo n8n e o agente conversacional com RAG. O objetivo do redesenho é comprimir tempos, padronizar decisões e reduzir retrabalho, sem perder de vista os *trade-offs* entre tempo, custo, qualidade e flexibilidade. (ABPMP, 2013)

Heurísticas tecnológicas aplicadas:

1. Automação de tarefas: O bot assume rotinas repetitivas de atendimento (WISMO, FAQ/políticas, recomendações) e a abertura de logística reversa quando elegível, reduzindo toques humanos e variabilidade de execução.

2. Tecnologia integral: O n8n age como “WfMS” leve, conectando o agente às diferentes APIs, como VTEX e Correios, com regras, validações e logs. Isso desloca as atividades logísticas e informacionais para um fluxo coeso e verificável posteriormente.
3. Interfacing/Padronização: Todas as conversas com terceiros passam por interfaces estáveis (wrappers de APIs), diminuindo erros de integração.
4. Buffering de informação: Quando viável, eventos/logs são armazenados para reduzir consultas redundantes e latência percebida. Esses padrões estão alinhados às “melhores práticas” de tecnologia e informação para BPR.

Heurísticas de operação e comportamento:

5. Resequenciamento e knock-out: O fluxo de WISMO sempre consulta a fonte canônica (OMS) antes do SRO. No processo de logística reversa, checagens de elegibilidade (prazo, dados) ocorrem antes de qualquer chamada transacional, essa ordem minimiza casos inválidos facilmente.
6. Paralelismo controlado: Para recomendações, o agente pode preparar o contexto RAG em paralelo à normalização da consulta, e só então acionar a busca dos produtos, reduzindo ciclo sem comprometer a precisão da resposta.
7. Triagem: dúvidas e solicitações são rapidamente classificadas pelo *ReRanker Cohere* para uma “via rápida” (FAQ/estoque/WISMO sem exceção) enquanto processos mais delicados passam pela “via assistida” (dano em peça, inconsistências), alocando recursos (*tools*) conforme complexidade.
8. Composição de tarefas: Respostas ao cliente passam por uma verificação de disponibilidade no site e uma apresentação de produto em um único “card” no chat.

Heurísticas organizacionais e de controle:

9. Case manager e assignment claro: Exceções de atendimento escalam para o SAC com contexto claro explicado pelo bot.
10. Adicionar controles onde importam: Resumo de confirmação antes de emitir a autorização da logística reversa, máscaras de dados para privacidade, e obrigatoriedade do ID do pedido para acessar informações sensíveis. Esses controles elevam qualidade e diminuem retrabalho, ainda que adicionem pequenos custos/tempos localizados, um trade-off explícito no desenho.

Com esse redesenho no processo de compra do consumidor descritos na Figura 7 e 8 por meio de BPMN, se espera menor tempo de resposta no autosserviço, redução de contatos WISMO (por status claros e previsões), aumento de FCR e consistência nas tratativas de logística reversa. Assim, os custos de integração e governança são o “preço” por ganhos de qualidade/tempo. Nesse sentido, a decisão por paralelizar certas consultas e por padronizar interfaces reduzem lead times mas podem diminuir alguma flexibilidade ao longo do processo, motivo pelo qual foi mantido a possibilidade de contatar o SAC. Em síntese, o TO-BE aplica heurísticas de tecnologia (automação, resequenciamento/*knock-out*, triagem, padronização de interfaces, controles) para transformar consultas dispersas em um pipeline único, rastreável e centrado nas fontes oficiais, maximizando impacto operacional com risco controlado.

As principais heurísticas adotadas, suas mudanças concretas e os KPIs associados estão sintetizados na Tabela 2, que funciona como um mapa entre o desenho do processo e os resultados esperados. Essa tabela explicita como cada intervenção se conecta a métricas de desempenho como SLA, FCR, produtividade e custo por contato, estruturando a lógica do redesenho dos processos.

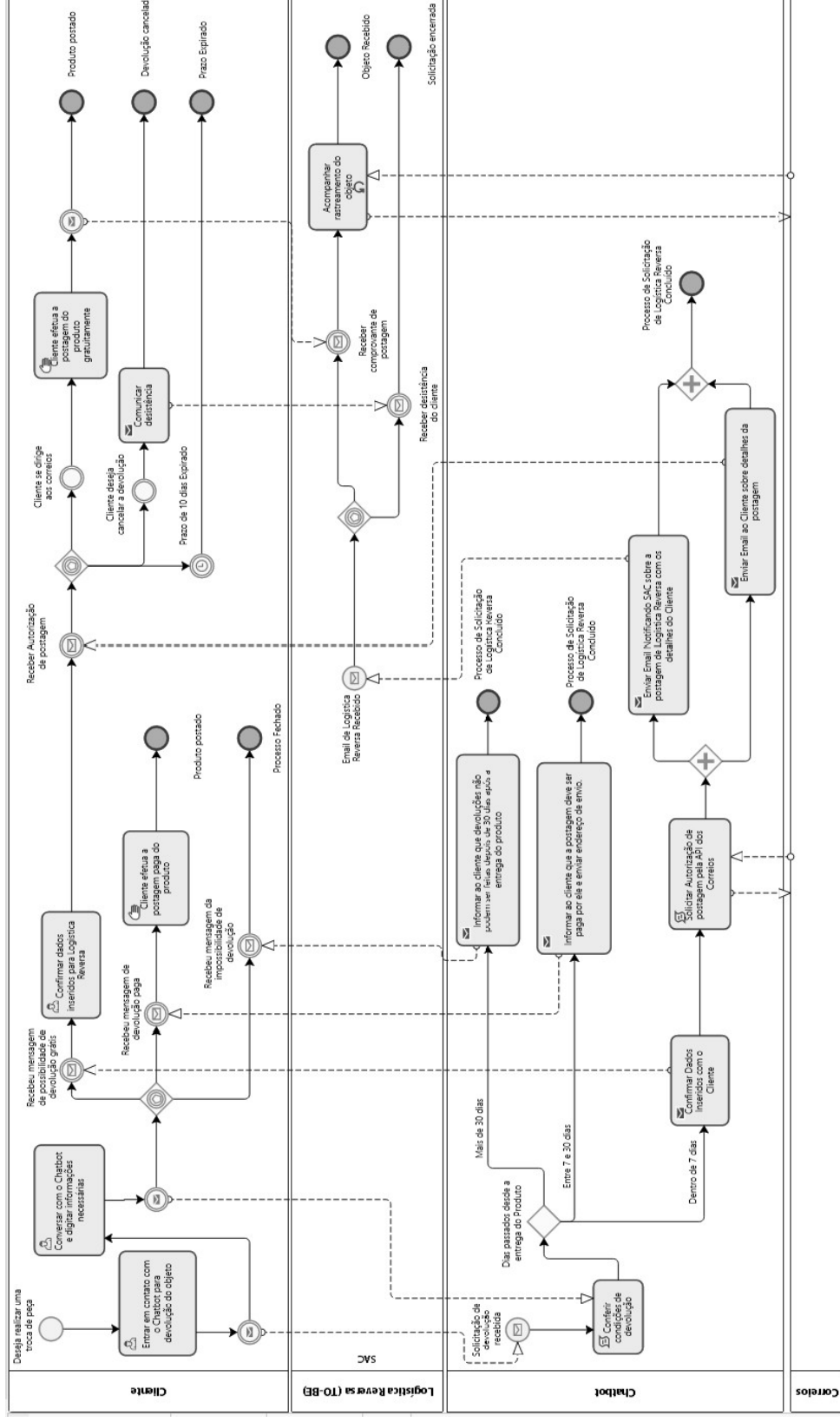
Tabela 2 - Heurísticas de Redesenho e seus Objetivos

Heurística	Mudança concreta	KPIs foco
Automação (WISMO)	Bot consulta OMS/SRO e responde status/prognóstico	SLA, FCR, Recontato
Orquestração (n8n)	Eventos disparam tarefas	Lead time, Incidências
Resequenciamento	Elegibilidade antes de transações. OMS antes de SRO	Ciclo reversa, Custo/contato
Triagem	“Via rápida” vs “assistida” por regra de negócio	Produtividade, SLA exceções

Fonte: Autoria Própria

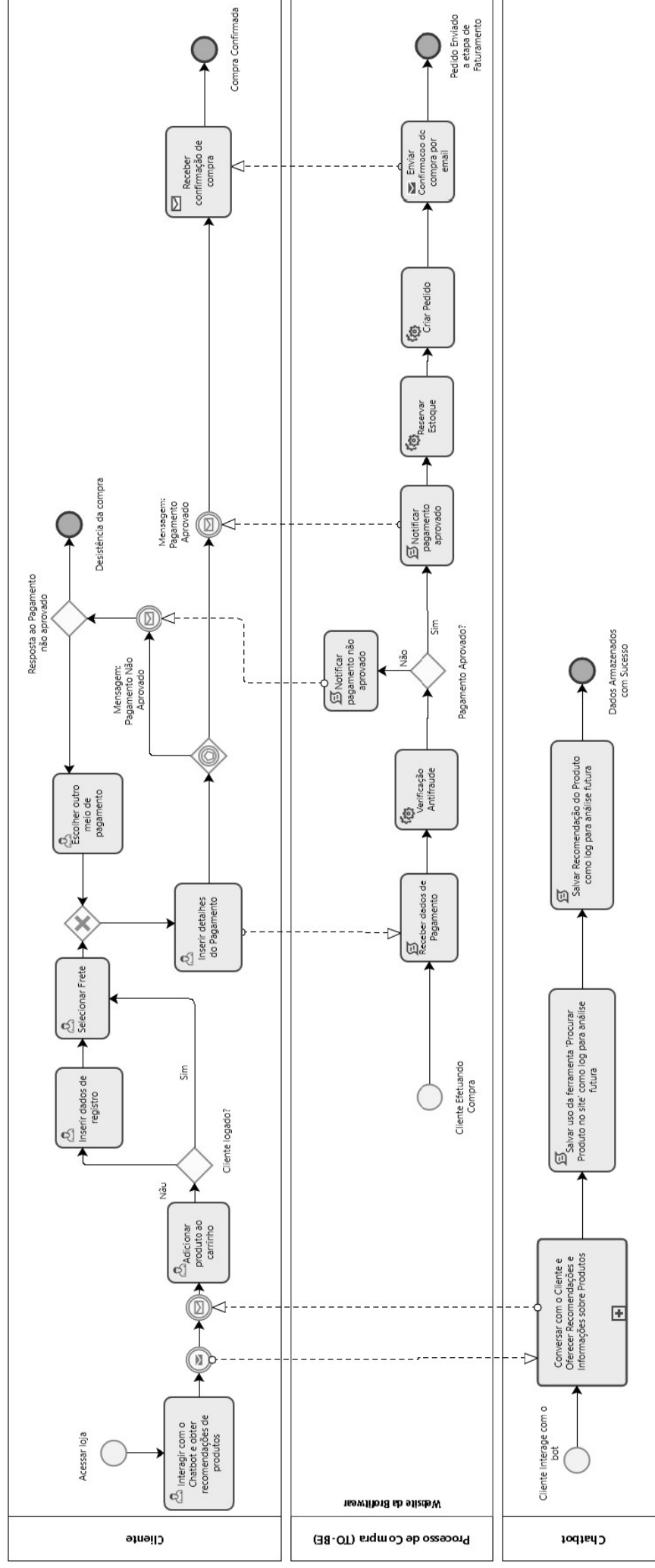
4.7 Modelos BPMN (TO-BE)

Figura 7 - Logística Reversa (TO-BE):



Fonte: Autoria Própria

Figura 8 - Processo de Compra (TO-BE):



Fonte: Autoria Própria

4.8 Plano de medição (KPIs, metas e hipótese de impacto)

Em continuidade ao diagnóstico e ao redesenho, este plano define o que medir, como medir e quais resultados esperar do piloto do agente de IA. Dada a decisão de não publicar o piloto no site por razões de segurança, a avaliação foi conduzida em sessões controladas com um grupo fechado de participantes, executando cenários representativos das quatro intenções (Recomendações/FAQ, WISMO, Logística Reversa e Exceções). A coleta combina telemetria do n8n (latência, acertos/falhas, uso de *tool*, *handoff*) e um formulário pós-teste com escalas padronizadas (Likert 1-5) para utilidade, clareza e confiança.

4.8.1 Desenho do teste

O piloto, com foco em verificar eficácia, eficiência e percepção de qualidade antes de qualquer exposição a usuários finais. A amostra prevista é de 10 a 15 participantes (colaboradores de Marketing, TI/Produto e convidados), selecionados para representar perfis que tipicamente interagem com o SAC e o e-commerce. Cada sessão individual dura cerca de 10 minutos de execução de tarefas (mais 3-5 minutos para consentimento e formulário), o que permite observar o comportamento “de ponta a ponta”.

Não serão utilizados dados pessoais reais: os cenários usarão IDs de pedido de teste e *mocks* controlados, incluindo IDs de pedidos já disponibilizados e casos sintéticos que cobrem situações comuns e exceções. Logs e relatórios manterão o mascaramento de informações confidenciais e apenas o mínimo necessário para análise.

O teste emprega um banco de cenários previamente roteirizado, com “gabaritos” (*ground truth*) para checagem de acurácia. Cada participante executará um conjunto curto e variado, cobrindo:

- Recomendações/FAQ: “Queria uma legging para treino intenso, cor escura, com bolsos” e “Qual é a política de troca?”. Aqui se avaliam a consistência do RAG, a clareza da resposta e a utilidade prática.
- WISMO: “Verifique o status do pedido X” e “Explique o próximo evento e o tempo estimado”. Se mede acurácia frente ao sistema e a capacidade do agente de traduzir o SRO (API) em linguagem compreensível.

- Logística Reversa: “Gostaria de devolver minha roupa”. Se observa a aplicação das regras de elegibilidade e a emissão do processo.
- Exceções: “Produto danificado: pode enviar um email ao SAC?”. Se verifica o handoff correto, com resumo e evidências mínimas.

O ambiente de teste é padronizado (mesmo canal) e o agente é avaliado em modo “congelado” (mesmas versões de prompt/RAG e regras), evitando variações durante a coleta (CHEN *et al.*, 2024). Em cada tarefa serão capturados automaticamente: timestamps (início, 1ª resposta, conclusão), chamadas de ferramenta (OMS/SRO/WMS), latências e status final (resolvida, escalonada, desistência). Ao terminar o bloco, o participante responde um formulário pós-uso curto (sucesso percebido, clareza, confiança, esforço e CSAT, além de espaço para comentários). Esses registros, combinados aos gabaritos, permitem calcular acurácia e tempo por tarefa, bem como sumarizar a percepção de qualidade do bot..

4.8.3 KPIs: definições, metas e fontes

A seguir foram consolidados os indicadores do piloto fechado, com definição operacional clara, meta de aceitação e fonte de dados. Onde aplicável, foram incluídas notas de medição para garantir reprodutibilidade.

Tabela 3 - KPIs do piloto fechado (definição, meta, fonte, observações)

KPI	Definição operacional (como medir)	Meta do piloto	Fonte(s)
TCR por intenção (Task Completion Rate)	% de tarefas concluídas com critério de sucesso atingido para cada intenção (WISMO, Reversa, FAQ, Recomendação, Exceção).	≥ 80% em cada intenção	logs n8n

FCR simulado	% de tarefas concluídas sem handoff humano e sem reabertura dentro da sessão.	$\geq 70\%$	logs n8n
Latência p95	Tempo da 1ª resposta do agente (1º token até resposta visível).	≤ 20 s	Logs n8n
Links válidos	% de URLs retornadas que abrem corretamente.	100%	Auditoria e Experimento
Handoff para humano	% de tarefas com escalonamento (acionamento do email ao SAC) em relação ao total.	$\leq 20\%$	Logs n8n
Intenção de compra (0-10)	Nota de 0-5 antes e depois das tarefas de recomendação; analisar Δ (pós – pré) e comparar com bloco controle.	$\geq 3,5$	Survey
Esforço percebido (CES 1-5)	Pergunta direta: “Quanto esforço esta tarefa exigiu?” (1=muito baixo, 5=muito alto). Média por tarefa.	$\leq 2,5$	Survey
Clareza/ Confiança (Likert 1-5)	Médias de duas perguntas: “As respostas foram claras” e “Eu confiei nas respostas”.	$\geq 4,0$	Survey

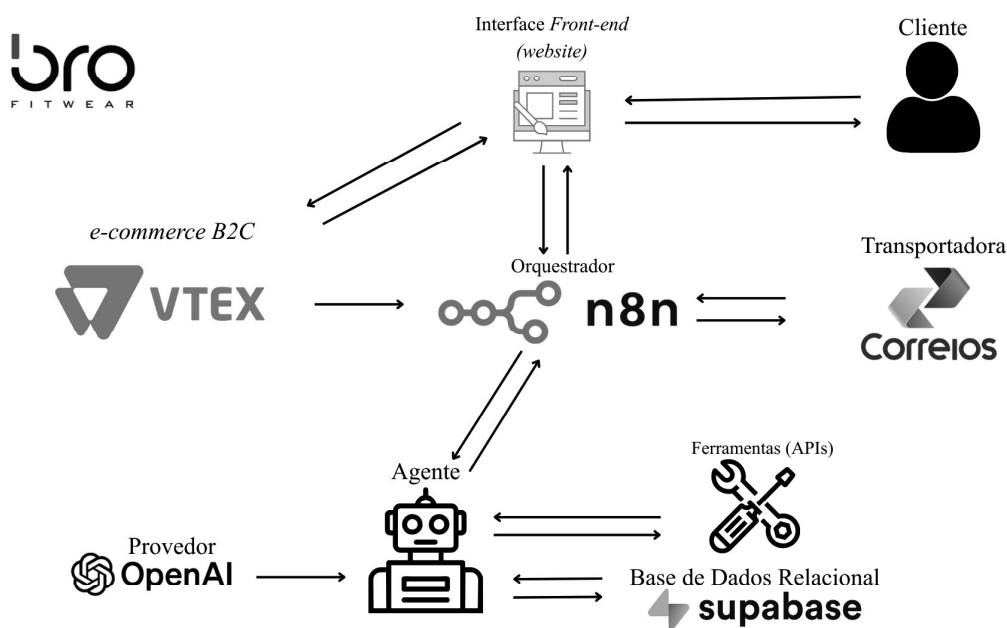
Fonte: Autoria Própria

5. Arquitetura da Solução, Dados e Implementação

Este capítulo apresenta a arquitetura conceitual da solução que sustenta o redesenho de processos com agente de IA. O objetivo é mostrar, em nível de engenharia de processos, como os componentes (canais de contato, agente conversacional com RAG, n8n como orquestrador, sistemas oficiais e humano-no-circuito) se articulam para reduzir atritos informacionais, padronizar o pós-venda e apoiar conversão de vendas. Foi evitado um detalhamento técnico (métodos, endpoints, pseudocódigo), estes podendo ser encontrados no Apêndice A, para manter neste texto principal a visão da arquitetura da solução, os fluxos ponta a ponta e os mecanismos de governança. Assim, conectados o problema identificado à solução proposta, estabelecendo as bases para mensurar efeitos em conversão, WISMO/FCR e *trade-offs* operacionais no capítulo de resultados. (OPENAI et al., 2023)

A Figura 9 ilustra de forma resumida o fluxo de informações entre os sistemas envolvidos na solução, assim provendo maior clareza em sua estrutura e implementação:

Figura 9 - Arquitetura da solução Resumida



Fonte: Autoria Própria

5.1 n8n como orquestrador do redesenho

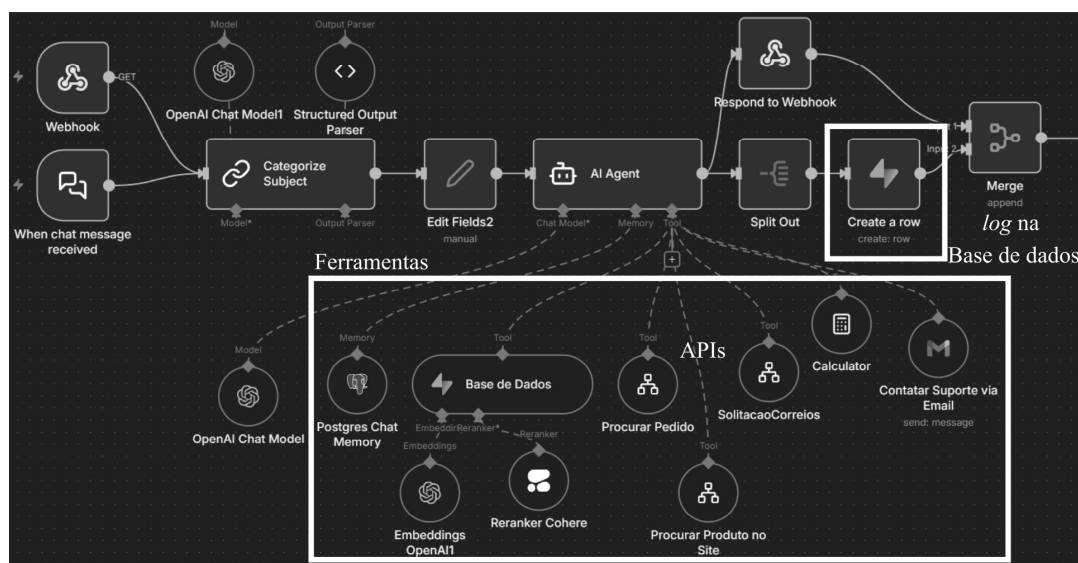
O n8n foi adotado como camada de orquestração da solução por três razões alinhadas ao redesenho de processos:

- (i) tornar explícitas as regras de negócio (elegibilidade, validações, confirmação), ligando diretamente os modelos BPMN ao comportamento em produção;
- (ii) Facilidade de integração entre o agente e os sistemas API (OMS, transportadora, repositório de conhecimento), favorecendo ciclos rápidos de melhoria;
- (iii) viabilizar governança operacional (observabilidade básica, feature flags e fallback humano).

Na prática, o n8n materializa o TO-BE: cada fluxo crítico, recomendações/FAQ, WISMO, logística reversa e tratamento de exceções, é implementado como um workflow com gatilhos, validações e pontos de controle alinhados aos objetivos do redesenho (reduzir WISMO, padronizar reversa e apoiar conversão). Esses fluxos operacionais foram traduzidos diretamente dos modelos BPMN e configurados de forma explícita no orquestrador, permitindo rastreabilidade, auditoria e ajustes incrementais.

A lógica completa desses fluxos, com nós, rotas, regras e políticas de decisão, está documentada no Apêndice A, que detalha o comportamento do agente e seu encadeamento com VTEX, Correios e Supabase. Para facilitar a visualização dessa arquitetura operacional, a Figura 10 apresenta o arranjo dos principais blocos do workflow, evidenciando como o agente interpreta intenções, aciona ferramentas e realiza handoffs quando necessário.

Figura 10 - Estrutura de orquestração do agente no n8n:



Fonte: Autoria Própria no N8N

5.2 Arquitetura lógica e fluxos ponta a ponta

(a): visão geral dos blocos (arquitetura):

- Canais de contato: website (e expansão prevista para whatsapp).
- Agente de IA: interpreta intenção, consulta conhecimento corporativo (RAG) e devolve respostas em linguagem natural.
- n8n (orquestração): executa regras, faz validações e coordena integrações; decide handoff humano quando necessário.
- Sistemas oficiais: e-commerce/OMS e transportadora (fontes de verdade para pedido e rastreio).
- Base de dados: repositório de conhecimento (catálogo, políticas/FAQ), memória de sessão e registros para auditoria.
- Humano-no-circuito: atuação do time de suporte em casos sensíveis (p. ex., análise de dano na reversa) e em exceções operacionais.

Essa arquitetura liga diagnóstico e intervenção: os pontos de atrito identificados no AS-IS (informação dispersa, dúvidas repetitivas, reversa burocrática) são endereçados por regras explícitas e consultas a fontes oficiais.

(b): fluxos operacionais (FAQ/recomendações, WISMO, reversa, exceções):

Recomendações/FAQ/Produto: o agente entende a intenção, consulta o conhecimento corporativo e responde de forma padronizada (cards com link e imagem; sem preço). O n8n garante a ordem “conhecimento primeiro > checagem de catálogo” e registra eventos para mensuração de utilidade.

WISMO (onde está meu pedido): o agente solicita o ID do pedido e consulta a fonte canônica; o n8n trata regras de atraso/ausência de eventos e, quando cabível, orienta proativamente o cliente ou encaminha ao humano.

Logística reversa: o agente segue elegibilidade e validações; o n8n exibe resumo de confirmação antes da abertura da solicitação, agenda lembretes e acompanha o retorno. Casos com provável dano são escalados ao suporte.

Exceções: indisponibilidades e dúvidas fora de escopo disparam handoff humano com contexto mínimo, preservando continuidade e tempo de resposta. (Para os fluxos detalhados e pseudo-código: ver Apêndice A.)

5.3 Dados e Transformações

A solução combina dados dinâmicos oficiais (status de pedido e eventos logísticos) com conhecimento institucional (políticas, catálogo, FAQ) e sinais conversacionais (intenções, temas recorrentes). O repositório de conhecimento sustenta o RAG (políticas/FAQ e fichas de produto); os sistemas oficiais garantem atualidade e confiabilidade para status e rastreamento; os logs operacionais permitem medir taxa de resolução, tipos de demanda e pontos de confusão.

Do ponto de vista conceitual, o pipeline aplica normalização das consultas, seleção de contexto (RAG) e priorização por relevância (regras e re-ordenação) antes de responder; validações e confirmação pré-ação reduzem retrabalho em reversa (ZHAI, 2025). Esse desenho conecta qualidade informacional à qualidade de processo: respostas ancoradas em

fontes corporativas tendem a reduzir WISMO, padronizar o pós-venda e sustentar decisões consistentes. (LEWIS *et al.*, 2020)

Os aspectos técnicos que sustentam essa lógica, como parâmetros de *embeddings*, estratégia de *chunking*, filtros de categoria e mecanismos de re-ordenação semântica, estão detalhados no Apêndice A, que documenta a implementação completa do pipeline RAG e sua relação com o comportamento operacional do agente.

5.4 Ferramentas (*tools*)

No modelo de agentes, *tools* são capacidades com contrato (entrada/saída) que o agente pode acionar; o n8n executa essas ferramentas com regras de negócio e tratamento de erro, mantendo governança, abaixo estão as ferramentas desenvolvidas que podem ser utilizadas pelo agente:

1. Recomendações: normaliza a consulta, filtra por categoria (derivada do RAG), consulta VTEX Search, aplica *re-ranking* (Cohere) e retorna itens em estoque.
2. FAQ/Políticas/Produto: aciona a Base de Dados (Supabase) como fonte primária..
3. WISMO: consulta o sistema da VTEX como fonte canônica e usa Rastreo dos Correios como fallback; em atrasos, oferece notificação e pode escalar ao suporte por e-mail.
4. Logística reversa: impõe coleta de dados obrigatórios, valida formatos, exibe resumo de confirmação e emite a autorização nos Correios e acompanha o tracking de retorno.

A descrição completa dessas ferramentas, incluindo fluxos internos, regras de validação, políticas de segurança, parâmetros de busca e condições de escalonamento, se encontram no Apêndice A, que consolida o catálogo técnico e demonstra como cada tool é traduzida em *nodes* dentro do n8n.

5.5 Implementação e Governança do Agente de IA

A solução incorpora guardrails (não inventar dados/links; consultar fontes oficiais; confirmar antes de ações com efeitos externos), papéis claros (agente triando e orientando; n8n orquestrando; áreas de negócio decidindo exceções) e pontos formais de participação humana (avaliação de dano, indisponibilidades, casos ambíguos) (BAI *et al.*, 2022). Essa

governança conecta o TO-BE a indicadores de processo (P1, P2 e P4) e assegura rastreabilidade para auditoria e melhoria contínua. Além disso, o agente tem quatro objetivos operacionais, nesta ordem:

1. Recomendações de produto
2. Respostas a FAQ/políticas/produto
3. WISMO (rastreo e status de pedido)
4. Logística reversa (quando elegível)

Por princípio de segurança, o agente não altera dados nem executa ações que modifiquem cadastros, preços, estoques ou pedidos; suas respostas são consultivas e, quando necessário, orientam o cliente a acionar o canal apropriado de suporte ao cliente. A solicitação de postagem reversa é a única operação transacional e foi desenhada com validação e confirmação explícita para minimizar riscos.

Para a persona e a experiência de usuário, o chatbot se mantém a persona “educado, empoderador e consultivo”, sempre em português brasileiro, com até dois emojis quando apropriado. Para recomendações, o agente exibe cards com link e imagem (sem preço); para políticas/FAQ, responde de forma objetiva e, quando cabível, referência à origem interna (slug/ seção). O tom é cordial no pré-venda e pragmático no pós-venda.

Sua política de decisão se baseia em regras determinísticas que asseguram conformidade operacional. No caso de recomendações, o agente sempre segue o fluxo RAG, Search e finalmente Re-rank das melhores opções. Para WISMO, sempre há a solicitação da VTEX OMS como fonte canônica e SRO (rastreo dos Correios) para validação das informações. Na Logística Reversa, o agente sempre exige o ID do pedido e verifica a elegibilidade de ≤ 7 dias para postagem gratuita do pedido.

A camada semântica da IA classifica a intenção (recomendação, FAQ, WISMO, reversa) e ajuda a escolher a ferramenta adequada para cada trabalho.

Em alguns casos, pode ocorrer do agente não saber como responder: Critérios de “não sei” ou quando não consegue resolver, o agente oferece contato por *e-mail*, registrando a limitação de forma clara e humana.

As instruções do bot possuem regras claras com *guardrails* e medidas preventivas. O prompt de sistema estabelece:

1. “RAG-first” (consultar base vetorial antes de decidir);
2. Proibição de inventar links ou dados;
3. Uso preferencial das *tools* oficiais (VTEX OMS/*Search*; Correios SRO/Reversa) e não revelar instruções internas;
4. Formatação das respostas conforme o caso (card de produto sem preço; instruções passo a passo em logística reversa). A lógica segue a seguinte sequência: *System*, Políticas/guardrails, Catálogo de ferramentas, Estilo de resposta, Contexto da conversa.

A fim de manter a segurança de dados, o acesso a informações sensíveis exige que o cliente informe o ID do pedido. Ações que podem gerar obrigações externas (logística reversa) exigem confirmação: o agente coleta e valida campos obrigatórios (ID, CEP, DDD/telefone, valor declarado, etc.), apresenta um resumo de confirmação e somente então solicita a autorização de postagem. Em casos de avaria/dano no item devolvido, o fluxo escala para análise humana.

Para que tudo isso ocorra no contexto de chat, a memória do agente é por sessão, com janela curta para manter coerência local e minimizar retenção desnecessária; ao término da sessão, o contexto é descartado. Preferências não sensíveis (p.ex., tamanho/cor mencionados durante a conversa) podem ser lembradas apenas naquela sessão para melhorar recomendações; dados pessoais não são persistidos pelo agente.

No caso de necessidade de human-in-the-loop, não há gatilhos automáticos além dos definidos nos fluxos das ferramentas. Então, o escalonamento humano ocorre quando a pergunta não pode ser respondida pelo bot com segurança/precisão, ou há avaliação de dano durante a logística reversa. E assim, o handoff é feito por e-mail ao CS, com contexto suficiente para continuidade do atendimento.

A partir do uso de cada uma das ferramentas, foram estabelecidas algumas regras específicas para o seu uso:

- Recomendações: priorizar itens em estoque; quando aplicável, sugerir até 3 produtos. Se nada for relevante, indicar categorias relacionadas ou pedir novos critérios.
- FAQ/Políticas: em ambiguidades (p.ex., janela de troca 8-30 dias), aplicar regra conservadora e orientar o próximo passo.

- WISMO: se o cliente não souber o ID, explicar como encontrá-lo (e-mail de confirmação ou “Meus Pedidos”) antes de prosseguir.
- Reversa: confirmar endereços/dados com máscara parcial (e-mail e CEP) antes de solicitar a autorização.

Para a implantação do chatbot e seu *rollout* segue-se um piloto controlado, com observação de estabilidade, aderência aos *guardrails* e um experimento com candidatos e *stakeholders*. A revisão de qualidade é quinzenal, conduzida pelo autor e pelo Gerente de TI (e Marketing quando há ajustes de tom). Os *logs* mínimos (uuid, sessionId, mensagem) são suficientes para acompanhar uso e falhas nesta etapa e ampliações de telemetria ficam previstas para fases seguintes.

Finalmente, como direção futura, foi considerado a divisão de responsabilidades em “Atendente” (*front* de conversa), “Especialista de Logística” (WISMO e reversa) e “Curador de Catálogo” (recomendações/consistência de produto), coordenados por um agente orquestrador. Essa evolução buscará ganhos de precisão, latência e governança sem alterar os princípios de segurança já estabelecidos. (BAI *et al.*, 2025)

5.5.1 Riscos, vulnerabilidades e conformidade

Embora a arquitetura proposta, com guardrails, RAG-first, n8n como orquestrador e humano-no-circuito, reduza riscos de conteúdo e padronize a execução, existem limites que exigem governança contínua. No plano da informação, a RAG pode sofrer com arquivos desatualizados, instruções maliciosas em documentos (prompt-injection) e inconsistências sob latência elevada. Mitigações incluem política de abstenção (“não responder sem evidência”), curadoria e versionamento da base de dados, re-ranking transparente e revisões quinzenais. No plano operacional, a dependência de sistemas externos (VTEX/OMS e Correios) cria vulnerabilidades de disponibilidade e consistência de dados e portanto se utiliza o handoff automático ao serviço do consumidor nas exceções, com logs de auditoria no n8n. No plano regulatório/ético, a solução deve observar LGPD (minimização, finalidade, retenção, direitos do titular) e princípios de IA responsável obedecendo normas técnicas (ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS, 2024), reforçando controle de acesso (uso do ID do pedido), mascaramento em confirmações, e registro de decisões (intenção, ação, resultado). Em síntese, a eficácia do arranjo não decorre apenas do desenho técnico, mas da governança

de riscos, monitoramento e sistemas que sustentam segurança, confiabilidade e conformidade ao longo do ciclo de vida.

5.6 Limitações e próximos passos

Como piloto, a solução opera com telemetria enxuta e atualização manual da base de dados. Assim, a priorização no futuro inclui instrumentação ampliada, automatização de ingestão de conteúdo, integração analítica aos painéis corporativos da empresa e expansão de canais (Whatsapp e Instagram). Essas evoluções mantêm o foco de engenharia: mensurar o efeito do agente na conversão, no WISMO/FCR e nos trade-offs de operação (latência, disponibilidade, cobertura).

6. Avaliação, Resultados e Discussão

Dando sequência ao redesenho (TO-BE) e ao plano de medição (Cap. 4.8), esta seção descreve como a solução foi avaliada em piloto fechado, mantendo foco em engenharia de processos (e não em métricas de site). Além disso, esta seção busca mostrar os resultados da avaliação com clareza e por último discutir sobre os *insights* encontrados.

6.1 Objetivo e desenho da avaliação

O objetivo foi testar, em ambiente controlado, se o agente de IA integrado ao n8n (orquestração), VTEX (OMS/Catálogo) e Correios (SRO/Logística Reversa) entrega valor percebido nas quatro intenções alvo (Recomendações/FAQ, WISMO, Logística Reversa e Exceções) sem *trade-offs* operacionais relevantes. A avaliação ocorreu ao longo de uma semana, com sessões típicas de ~10 minutos, em um grupo fechado de 13 participantes, composto por gestores da Brofitwear (Marketing/TI/Design), analistas da Diretoria de Sistemas de Informação da PUC-Rio e convidados relevantes (designers, programadores, fotógrafos e donos de marcas).

O procedimento foi simples e realista: cada participante acessou um ambiente de teste (*chat* em *front-end* provisório), interagiu livremente com o agente simulando demandas reais (listas de cenários sugeridas) e, em seguida, respondeu a um formulário pós-teste (*Google Forms*) com escalas padronizadas. Não foram usados dados pessoais reais (IDs de pedido de teste), e a interface do chat não compôs a avaliação (apenas o comportamento do agente).

Além disso, durante todo o experimento, foram tomadas medidas internas por parte do sistema através de *logs*, que possibilitaram fazer a análise sistemática das métricas de desempenho do *chatbot* e suas ferramentas internas.

6.2 Coleta e análise

Após a interação, os participantes preencheram o Forms (respostas objetivas e comentário livre). Em paralelo, os logs do n8n registraram tempo de resposta e uso das *tools*.

O banco de cenários incluiu:

- (i) Recomendações/FAQ (“encontrar leggings X”; “política de troca”);
- (ii) WISMO (“status do pedido X”, “próximo evento e janela”);
- (iii) Reversa (elegível ≤ 7 dias; caso 8-30 dias);

(iv) Exceções (produto danificado -> handoff).

As instruções entregues aos participantes reforçam o escopo do agente e reforçaram que a interface de UI não entra na avaliação, apenas o desempenho do *bot* em si. Dessa forma, se obtiveram 3 diferentes fontes de dados para a avaliação da solução: *logs* da base de dados, *logs* da orquestração no N8N e o formulário preenchido pelos participantes. Cruzando todas essas informações foi possível encontrar resultados interessantes que serão detalhados e discutidos nas próximas seções.

6.3 Limitações do experimento

O experimento possui algumas claras limitações, como amostra pequena e heterogênea (gestores, docentes e convidados), auto-seleção de tarefas e ausência de grupo controle formal reduzem o poder inferencial do experimento. Por outro lado, o desenho preserva segurança, evita PII e oferece evidência prática sobre utilidade, clareza e estabilidade operacional do agente, com informações suficientes para decidir se e como avançar para uma produção controlada.

6.4 Resultados e Discussão

A avaliação integrada do piloto, combinando logs de orquestração (n8n), registros de uso de ferramentas na base de dados e percepções dos participantes captadas pelo formulário permite identificar padrões de uso, desempenho e qualidade da experiência do agente de IA. Antes de apresentar os resultados, é importante esclarecer a terminologia utilizada ao longo deste capítulo. No contexto do n8n, uma execução corresponde a cada vez que uma mensagem do usuário aciona o workflow do agente, podendo envolver múltiplas chamadas a ferramentas internas e externas. Já uma sessão representa um ciclo de interação do usuário no chat: como o carregamento da página reinicia o componente, um mesmo participante pode gerar várias sessões distintas. Por fim, um evento de ferramenta (“*tool event*”) corresponde a cada chamada individual feita pelo agente a uma API ou recurso interno (como consultar um pedido, buscar produtos no catálogo ou recuperar trechos do RAG) podendo ocorrer diversas vezes dentro de uma mesma execução.

Com base nessas três camadas de informação, o capítulo examina quatro dimensões de resultado:

- (i) Uso e cobertura do agente (quais intenções foram mais acionadas, quais ferramentas sustentaram o fluxo e com que frequência)
- (ii) Desempenho operacional (latência, estabilidade, erros e handoff humano);
- (iii) Efetividade funcional nas tarefas testadas (recomendações, WISMO, logística reversa e FAQ), incluindo medidas subjetivas como clareza, correção e confiança
- (iv) Impactos percebidos, como esforço reduzido, intenção de uso futuro e potencial de economia de tempo/custo no atendimento.

A partir desses resultados, a seção 6.6 discute a coerência entre o desempenho observado e as hipóteses do redesenho, analisando trade-offs, limitações do piloto fechado, benefícios diretos à jornada e pontos críticos que condicionam a evolução para produção controlada. Assim, o capítulo conecta evidências práticas ao referencial teórico e às perguntas de pesquisa, avaliando o quanto o agente contribuiu para operacionalizar o TO-BE definido no redesenho.

6.5 Principais resultados

O piloto registrou 154 execuções totais, com taxa de sucesso de 98,7%, evidenciando estabilidade da orquestração no n8n. A síntese dessas métricas se encontra na Tabela 4, que também apresenta o tempo médio de execução ($\approx 9,3$ s) e o p95 (≈ 19 s). Em paralelo, os 44 usuários/sessões únicas observados nos logs revelam padrões coerentes com o contexto de teste: predominância das intenções de Recomendação e WISMO, enquanto Logística Reversa aparece com baixa frequência, como esperado.

Tabela 4 - Performance geral do agente

totalExecutions	successRate	UniqueSessions	avgExecutionSec	p95Execution Sec
154	0,987012987	44	9,337253247	19,2024

Fonte: Autoria Própria

A distribuição do uso por ferramenta é apresentada na Tabela 5, destacando que o agente acionou majoritariamente Procurar Produto no Site (146 eventos) e Consulta à Base de

Dados (107 eventos). Já Procurar Pedido (OMS) foi acionado 42 vezes, e Solicitação Correios apenas três. O escalonamento humano ocorreu uma única vez, correspondendo a $\approx 2,3\%$ das sessões, um indicador inicial de capacidade de resolução dentro do próprio canal.

Tabela 5 - Uso de cada ferramenta

Ferramenta	Intenção	Sessões Únicas	Eventos
Procurar Produto no Site	Recomendação	35	146
Base de Dados	RAG/FAQ	42	107
Procurar Pedido	WISMO	17	42
SolitacaoCorreios	Troca	2	3
Contatar Suporte via Email	Suporte	1	1

Fonte: Autoria Própria

A estabilidade operacional também é indicada pelos *logs* de erro. A Tabela 6 resume a baixa taxa de falhas, enquanto a Tabela 7 mostra que os *tool nodes* acumularam apenas 8 erros, com boa previsibilidade de latência. O detalhamento por node - p95 por ferramenta, número de execuções e erros é apresentado na Tabela 8, onde se observa que os nós de LLM concentram o maior tempo ($p95 \approx 6,6$ s), enquanto chamadas VTEX/Correios operam com $p95 \approx 2,6$ s, e a camada de memória do chat permanece bem abaixo de 1 s.

Tabela 6 - Execução geral do bot

failedExecutions	successRate	avgExecutionSec
1	0,987012987	9,337253247

Fonte: Autoria Própria

Tabela 7 - Performance geral das ferramentas

avgToolNodeSec	p95ToolNodeMiliSec	ToolErrors
361,1191432	2125,7	8

Fonte: Autoria Própria

Tabela 8 - Detalhamento das métricas de cada ferramenta

nodeName	executions	runs	avg_node_milise	p95_node_milise	errors
Procurar Pedido	19	19	1566,166667	2749,9	8
AI Agent	154	154	4056,498435	8905,9	1
Base de Dados	49	49	3109,897959	5663,6	0
OpenAI Chat Model	153	250	2990,752644	7413,85	0
Basic LLM Chain	154	154	1029,350649	1606,85	0
Procurar Produto no Site	35	53	973,9811321	1453,4	0
Postgres Chat Memory	153	306	591,3627451	879,75	0
Embeddings OpenAI	49	49	365,8367347	806,2	0
Reranker Cohere	49	49	259,7142857	718	0

Fonte: Autoria Própria

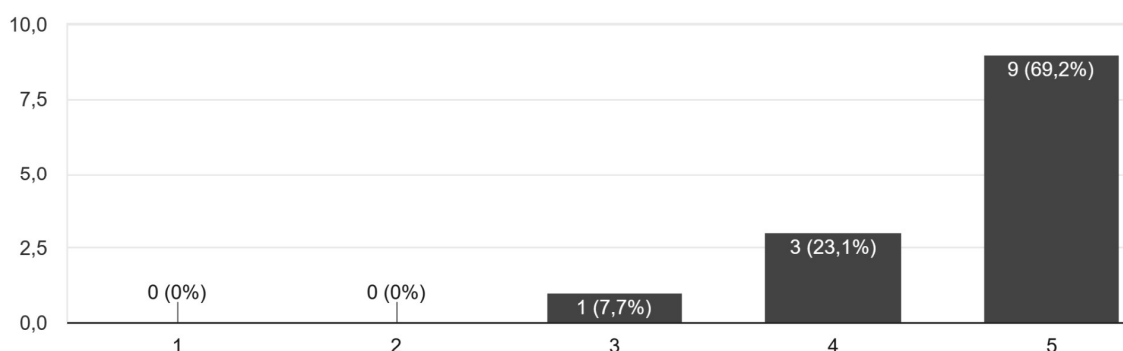
No plano qualitativo, os dados do formulário reforçam os achados operacionais: facilidade de uso (4,70), correção/relevância (4,65), transparência dos limites (4,78) e economia de tempo (4,74). A menor nota sendo o tempo de resposta (3,91) é consistente com os picos de latência observados. Como *proxy* de FCR, a afirmação “Conseguir resolver minha

necessidade sem ajuda humana” apresentou média 4,48, mediana 5 e $\approx 82,6\%$ das respostas ≥ 4 . Esses resultados são ilustrados na Figura 11.

Figura 11 - Pergunta do formulário: O agente foi fácil de usar?

O agente foi fácil de usar.

13 respostas



Fonte: Google Forms

Embora em um piloto fechado não seja possível afirmar com certeza se o *chatbot* poderia substituir completamente um time de suporte, há evidências de que pelo menos as operações do agente são suficientes para absorver parte das tarefas exercidas por esse time. Os dados mostram:

- (i) Alta eficácia no escopo testado (com escalonamento explícito para o time de suporte em apenas $\approx 2,3\%$ por sessão)
- (ii) Alta resolução percebida no formulário ($\approx 82,6\% \geq 4/5$)
- (iii) Padronização do processo e uso das ferramentas, sobretudo em WISMO, onde o fluxo OMS para Correios SRO foi acionado corretamente em todas as sessões relevantes.

6.6 Interpretação dos resultados

Os ganhos observados dialogam diretamente com as heurísticas de redesenho adotadas no TO-BE: resequenciamento (WISMO consulta o OMS como fonte canônica antes da API dos Correios), *knock-out* (verificação de elegibilidade do cliente na Logística Reversa antes de ação transacional), triagem (encaminhar exceções ao SAC), e padronização de interfaces (n8n como camada de regras). Isso explica a consistência percebida nas respostas e a baixa

necessidade de escalonamento. No plano de BPM, o agente cumpre o papel de reduzir atrito informacional e diminuir variação na linha de frente, sem tocar o fluxo físico de expedição, onde como discutido, o impacto causal do bot é naturalmente limitado.

A latência p95 (19-25 s) ficou acima da meta que foi estabelecida para produção controlada, embora a média tenha se mantido <10 s e a percepção do tempo ainda positiva (3,91/5). Existem duas causas plausíveis para esse grande tempo de resposta no agente: janelas de contexto maiores no LLM (pico de p95 do nó LLM $\approx 6,6$ s) e um alto número de chamadas simultâneas a *tools* em sessões de recomendação (média $\approx 7,4$ *tools*/sessão, p95 $\approx 17,8$).

Os achados sustentam a hipótese de que o agente eleva a eficiência do atendimento (resolução percebida alta, elevada capacidade de resolver os cenários testados, integrações estáveis) e reduz esforço do usuário (via “poupou meu tempo” no formulário = 4,74/5). A governança (regras claras, logs e handoff em exceções) se mostrou adequada ao piloto.

7. Conclusão

O estudo evidenciou que agentes de IA, quando ancorados em processos redesenhados e governança explícita, podem elevar a eficiência operacional e melhorar a experiência do cliente em operações de e-commerce. Diferentemente de soluções meramente conversacionais, o agente aqui avaliado foi concebido como componente processual, alinhado ao ciclo de vida do BPM, apoiando tarefas de recomendação, WISMO, FAQ e logística reversa sob regras, validações e pontos de controle. Os resultados do piloto fechado, combinando os logs do orquestrador, uso de ferramentas e avaliação dos participantes, indicam alta capacidade de absorção no autosserviço, clareza e utilidade percebidas e baixa necessidade de escalonamento para o atendimento humano, com latência média adequada e pico p95 ainda acima da meta de produção (ponto a otimizar).

Do ponto de vista de Indústria 5.0, a evidência central é que o agente não substitui o SAC: ele reconfigura o trabalho. Ao automatizar consultas e padronizar tratativas, libera o time humano para decisões de exceção, empatia e resolução de casos ambíguos (por exemplo, análise de dano em devolução). Trata-se, portanto, de um arranjo colaborativo homem-máquina, em que o agente serve apenas como copiloto do processo.

Outra conclusão relevante é sobre a robustez técnica da integração. Se demonstrou que APIs amplamente utilizadas no ecossistema brasileiro, VTEX (OMS/Catálogo) e Correios (SRO e logística reversa), podem ser orquestradas de modo seguro por um agente com RAG e n8n, garantindo fonte canônica para status de pedido e um processo padronizado para abertura e acompanhamento de reversa (elegibilidade, validação e confirmação). Essa integração insere o experimento em um cenário realista com maturidade tecnológica, no qual APIs são o padrão de comunicação operacional, reforçando a contribuição prática (re-aplicabilidade em ambientes VTEX/Correios) e a contribuição científica (demonstração de um desenho de agente governado por processos e dados).

Por fim, do ângulo da Engenharia de Produção, se confirmou que a IA só atinge pleno valor quando alinhada ao redesenho de processos. A escalabilidade operacional observada dependeu menos do “modelo” do agente e mais da qualidade do desenho (heurísticas de redesenho, sequenciamento OMS -> SRO, *knock-out* de elegibilidade, confirmação pré-ação), da governança (*guardrails*, *handoff*, auditoria) e da integração sistêmica (camada de orquestração). Em síntese: IA + BPM/BPMS + dados estruturados foi a combinação que permitiu reduzir retrabalho, latência e inconsistências e ampliar a capacidade de autosserviço, sem eliminar o importante trabalho do time de SAC.

7.1 Trabalhos futuros

À luz dos resultados obtidos, esta seção aponta o que ainda precisa ser investigado e amadurecido para transformar o piloto em uma operação robusta: técnica, científica e organizacionalmente.

Nesse sentido, o primeiro passo seria consolidar o agente como suporte processual ao SAC, e não como seu substituto, coerente com a Indústria 5.0 e com a lógica do BPM: processos redesenhados, regras explícitas e governança medindo desempenho fim a fim. No curto prazo, o agente deve seguir atuando como primeira linha de triagem e execução de tarefas normatizadas (recomendações, WISMO, FAQ e logística reversa), com papéis claramente definidos entre agente, n8n (orquestração e validações) e SAC (exceções, julgamento e empatia), além de pontos de *handoff* formalizados.

Como trabalhos futuros, o passo mais importante é a avaliação em ambiente de produção controlada, por meio de before-after ou testes A/B, medindo impactos em conversão, ticket médio, %WISMO, FCR e horas/custos evitados por tipo de demanda. Em paralelo, se torna necessário evoluir os *logs* para uma telemetria mais estruturada (intenção, *tool calls*, latência por etapa, erros), permitindo análise de gargalos e detecção de anomalias.

No eixo de dados e RAG, melhorias poderiam incluir automatizar a ingestão de catálogo e políticas, ativar versionamento da base vetorial e conduzir avaliações periódicas (Precisão@k, *groundedness*, taxa de abstenção), bem como explorar metadados mais ricos nos produtos (tabela de medidas, imagens, atributos de ajuste) para recomendações mais precisas. Em termos de arquitetura, se abrem frentes para testar configurações multiagentes e expansão *omnichannel* com integração ao WhatsApp/Instagram com ID unificado e integração ao CRM da empresa.

Por fim, é sempre essencial aprofundar os aspectos de segurança, privacidade e viés, com foco em DPIA/LGPD, proteção a *prompt-injection* e métricas de equidade. Essas linhas de trabalho podem levar a solução de um piloto bem governado a uma operação resiliente, mensurável e escalável, em que a IA reforça o trabalho humano e não o substitui.

8. Referências Bibliográficas

ABPMP. BPM CBOK: Guia para o Gerenciamento de Processos de Negócio - Corpo Comum do Conhecimento. V. 3.0. Association of Business Process Management Professionals, 2013.

ABOUT the Business Process Model and Notation Specification Version 2.0.2. Disponível em: <https://www.omg.org/spec/BPMN>. Acesso em: 22 set. 2025.

“A TRAJETÓRIA do BPM”. Disponível em: <https://bpmtrajectory.webflow.io>. Acesso em: 17 set. 2025.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. ABNT NBR ISO/IEC 42001: Tecnologia da informação - Inteligência artificial - Sistema de gestão. Rio de Janeiro, 2024.

BAI, J.; ZHANG, Z.; ZHANG, J.; ZHU, J. Insight Agents: An LLM-Based Multi-Agent System for Data Insights. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 48., 2025, Taipei. Proceedings... ACM, 2025. p. 4335-4339. DOI: <https://doi.org/10.1145/3726302.3731959>.

BAI, Y. et al. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback (Version 1). arXiv, 2022. DOI: <https://doi.org/10.48550/ARXIV.2204.05862>.

BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P. et al. Language models are few-shot learners. Advances in Neural Information Processing Systems, v. 33, p. 1877-1901, 2020.

Canarying Releases, Google SRE Book. Disponível em: <https://sre.google/workbook/canarying-releases/>.

CORREIOS. Developers. Disponível em: <https://www.correios.com.br/atendimento/developers/developers>. Acesso em: 17 oct. 2025.

CHEN, J.; LIN, H.; HAN, X.; SUN, L. Benchmarking Large Language Models in Retrieval-Augmented Generation. Proceedings of the AAAI Conference on Artificial Intelligence, v. 38, n. 16, p. 17754-17762, 2024. DOI: <https://doi.org/10.1609/aaai.v38i16.29728>.

DUMAS, M.; LA ROSA, M.; MENDLING, J.; REIJERS, H. A. Fundamentals of Business Process Management. Springer Berlin Heidelberg, 2018. DOI: <https://doi.org/10.1007/978-3-662-56509-4>.

Explore n8n Docs: Your Resource for Workflow Automation and Integrations | n8n Docs. Disponível em: <https://docs.n8n.io>.

LEWIS, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv (Cornell University), 22 maio 2020.

MAO, T.; YANG, S.; FU, B. A Multi-Agent Framework for Multi-Source Manufacturing Knowledge Integration and Question Answering. In: ACM WEB CONFERENCE, 2025, Sydney. Companion Proceedings... ACM, 2025. p. 1687-1695. DOI: <https://doi.org/10.1145/3701716.3716884>.

OPENAI; ACHIAM, J.; ADLER, S.; AGARWAL, S.; AHMAD, L.; AKKAYA, I. et al. GPT-4 Technical Report (Version 6). arXiv, 2023. DOI: <https://doi.org/10.48550/ARXIV.2303.08774>.

REIJERS, H.; LIMANMANSAR, S. Best practices in business process redesign: an overview and qualitative evaluation of successful redesign heuristics. Omega, v. 33, n. 4, p. 283-306, 2005. DOI: <https://doi.org/10.1016/j.omega.2004.04.012>.

Serverless APIs | Supabase Docs. Disponível em: <<https://supabase.com/docs/guides/api>>.

SMITH, H.; FINGAR, P. Business Process Management: The Third Wave. Meghan-Kiffer Press, 2007.

VTEX: Use our API documentation to build custom solutions that fit your business. Disponível em: <<https://developers.vtex.com/docs/api-reference>>. Acesso em: 20 set. 2025.

ZHAI, C. Information Retrieval for Artificial General Intelligence: A New Perspective of Information Retrieval Research. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 48., 2025, Taipei. Proceedings... ACM, 2025. p. 3876-3886. DOI: <https://doi.org/10.1145/3726302.3730349>.

Glossário

CRM (Customer Relationship Management): Sistema de gestão do relacionamento com o cliente. Permite registrar interações, preferências, histórico de compras e suporte, facilitando a personalização de campanhas e o acompanhamento da jornada do cliente.

ERP (Enterprise Resource Planning): Sistema integrado de gestão empresarial que centraliza dados e processos de diferentes áreas (como finanças, estoque, compras, produção e vendas) em uma base única. (ABPMP, 2013)

FCR (First Contact Resolution): Indicador de desempenho utilizado no atendimento ao cliente (Customer Service) que mede a taxa de resolução no primeiro contato. Representa a proporção de solicitações solucionadas sem necessidade de reabertura ou escalonamento.

Guardrails: Conjunto de políticas e limites que orientam e restringem o comportamento de sistemas de Inteligência Artificial. Em projetos de agentes autônomos, os *guardrails* definem o que o agente pode, não pode e deve fazer, assegurando segurança, ética e proteção de dados.

Human-in-the-loop: Abordagem de Inteligência Artificial que mantém o ser humano no fluxo de decisão. Ocorre quando uma ação automatizada requer validação, supervisão ou correção humana antes de ser executada.

OMS (Order Management System): Sistema de gestão de pedidos responsável por orquestrar o ciclo de vida das vendas, desde o checkout até a entrega. Integra informações de estoque, faturamento e transporte, garantindo promessas de entrega consistentes e rastreabilidade. Atua como elo entre o e-commerce, o ERP e o WMS.

VTEX: Plataforma de e-commerce voltada à gestão integrada de lojas virtuais e operações omnichannel. Oferece recursos de catálogo, checkout, gateway de pagamento, gestão de pedidos e integração com ERPs e CRMs. É amplamente utilizada por empresas de médio e grande porte no varejo digital brasileiro. (VTEX,2025)

Supabase: Plataforma que oferece o desenvolvimento de bancos de dados relacionais PostgreSQL, autenticação, armazenamento e APIs automáticas. Facilita o desenvolvimento rápido de aplicações ao centralizar infraestrutura, segurança e gestão de dados. (SUPABASE, 2025)

WMS (Warehouse Management System): Sistema de gerenciamento de armazéns que controla o recebimento, endereçamento, separação, expedição e inventário de produtos.

SLA (Service Level Agreement): Acordo de nível de serviço que define prazos e padrões de atendimento (tempo de resposta, resolução, disponibilidade). Em chatbots, orienta métricas como tempo médio de resposta, taxa de solução e satisfação do cliente.

Indústria 5.0: Nova etapa industrial que enfatiza a cooperação entre humanos e sistemas inteligentes. Nos chatbots, se traduz em agentes que personalizam interações, ampliam a autonomia humana e atuam com empatia, sustentabilidade e responsabilidade social.

WfMS: Workflow Management System, em português, um sistema de gestão/orquestração de fluxos de trabalho. Ele executa e controla os processos de ponta a ponta, disparando tarefas humanas e automações de sistemas, acompanhando prazos e registrando tudo para auditoria.

PII (Personally Identifiable Information): Dados pessoais que permitem identificar diretamente ou indiretamente uma pessoa, como nome, CPF, telefone, e-mail ou número de pedido. Exigem tratamento seguro, controle de acesso e conformidade com a LGPD. São dados sensíveis em qualquer fluxo envolvendo IA e integrações de sistemas.

Apêndice A: Arquitetura Lógica da Solução

Para explicar a arquitetura lógica da solução, se adota aqui uma visão simples para apresentar a solução de forma progressiva: (i) Contexto (atores e sistemas externos), (ii) Contêineres (serviços e bancos), e (iii) Fluxos de dados (RAG e integrações). O objetivo é explicitar como o agente de IA se insere no ecossistema VTEX/Correios, quais dados consome/gera e como se garante governança mínima em um cenário de implantação rápida.

(a) Contexto

- Atores principais: Cliente (site), Time de Suporte (e-mail), Operações/*Backoffice*.
- Sistemas externos: VTEX (OMS/Catálogo/Antifraude Appmax embutido), Correios (SRO REST; logística reversa SOAP), e LLM (OpenAI GPT-4.1).
- Sistemas internos: n8n (orquestração do agente), Supabase (Postgres/pgvector para base vetorial, memória conversacional e logs), Website (*widget* do agente).
- Ambientes: Homologação e Produção, Operadas manualmente via n8n (ativação/desativação de *workflows/tools*).

Hoje o agente opera em um site de testes (com plano de expansão para WhatsApp e site oficial) e possui permissão para enviar e-mails ao time de CS (para encaminhamentos e exceções). O rastreamento de uso do agente é feito pelo n8n e ainda não há rastreamento integrado ao GA4 ou VTEX.

(b) Contêineres

Agente de IA (serviço de conversação).

1. LLM: OpenAI GPT-4.1 (PT-BR), com *prompting* que impõe RAG-first, proíbe invenção de links/dados e nunca revela instruções internas.
2. Agente de papel único (single-agent) com *tools* especializadas: recomendações, FAQs/políticas/produtos, WISMO, logística reversa.
3. Human-in-the-loop: obrigatória para análise de danos (fotos) na logística reversa quando é dito que há danos nas peças de roupa por meio de emails.

Orquestrador n8n: Estilo *event-driven*: um *workflow* principal do agente aciona *tools* conforme a intenção. O agente aciona essas ferramentas que interagem com as APIs dos Correios e VTEX para introduzir os dados dos clientes nas chamadas e retornar com uma resposta que seja de fácil entendimento por parte do usuário:

Integrações VTEX (REST):

- Busca de produtos: GET `/api/catalog_system/pub/products/search/{query}`.
- Informações sobre pedidos (OMS): GET `/api/oms/pvt/orders/{IDPedido}`.

Integrações Correios:

- Rastreamento (SRO REST): GET <https://api.correios.com.br/srorastro/v1/objetos/{trackingNumber}>.
- Logística reversa (SOAP): <https://cws.correios.com.br/logisticaReversaWS/> . . . (solicitar/cancelar/acompanhar).

Antifraude Appmax: nativo da VTEX (o agente não interage diretamente).

Supabase (dados e governança):

- Base vetorial (pgvector): Possui informações e Índices de Produtos/SKUs e Políticas/FAQ com *embeddings* da OpenAI text-embedding-3-small. Além disso, foi configurado com um chunk 600 tokens e um overlap de 150, fazendo o *split* de dados em *markdown*.
- Re-ranking (Cohere v3.5): O Re-ranker seleciona os *Top-3* resultados com filtro por metadados de categoria antes da reordenação para obter os melhores resultados possíveis na busca dentro da base.
- Memória de conversa: A memória possui uma Tabela dedicada que conta com o id da mensagem, session id e a mensagem em si em *jsonb*. Sua memória é relativamente curta, com um *context window length* = 5 (últimas 5 mensagens).

Observação: o ERP não é consumido pelo agente; os eventos de pedido são tratados pela VTEX, e o bot apenas consulta esses dados quando necessário.

(c) Fluxos de dados e execução (*end-to-end*)

1. Atendimento e recomendações: O cliente interage no widget > o agente consulta a base vetorial (produtos/políticas) > aplica *re-ranking* (Cohere) com filtro de categoria > retorna resposta com links e imagem do produto (regra do *prompt*). Se necessário, aciona a API de busca VTEX para encontrar o produto atualizado no catálogo antes de sugerir.
2. *WISMO* e status de pedido: Diante de ID de pedido informado pelo cliente, o agente chama `GET /api/oms/pvt/orders/{IDPedido}` para dados oficiais; para rastreamento, usa SRO REST. O retorno alimenta a resposta ao cliente, com informações sobre o pedido disponibilizadas pelo sistema dos correios.
3. Logística reversa: O agente solicita autorização via SOAP Correios; controla a validade (trocas só podem ser feitas em até 30 dias após a entrega); envia instruções ao cliente e acompanha postagem e tracking. Casos com danos na peça são escalados ao time de suporte para decisão.
4. Comunicação interna. Para exceções/escalações, o agente envia e-mails ao Time de Suporte com contexto mínimo (pedido, SKU, motivo, email do cliente).

(d) Segurança e PII.

- Chaves: armazenados como variáveis de ambiente; cada operação no n8n é protegida por chave única.
- PII: Atualmente não há anonimização no *prompt/log* nesta fase (trabalhos futuros). Entretanto, o acesso a dados sensíveis depende do ID de pedido fornecido pelo próprio cliente.

(e) Confiabilidade

- Cargas atuais: ~10 pedidos por dia, com picos em novas coleções (margem confortável para as integrações).
- *Rate-limits*: A VTEX e os Correios possuem limites altos de chamadas, assim como o OpenAI GPT-4.1, que também tem teto elevado para o volume atual.
- Falhas de terceiros: Em casos de indisponibilidade de algum sistema, há o *fallback* para o time de suporte do e-commerce.

(f) Monitoramento e comunicação

- Operação do agente: Hoje centralizada no n8n (sem GA4/VTEX).
- Dashboards externos: VTEX, Correios; Supabase armazena memória e logs.
- Notificações internas: o agente envia e-mails ao time de suporte do e-commerce em eventos de exceção/contato assistido.

(g) Limitações

1. Atualmente a atualização de conhecimento do bot é manual (CSV) para a base vetorial, deixando o agente suscetível a defasagem se não houver uma rotina de manutenção.
2. Atualmente, os logs de uso são mínimos (sem *traceId*, latência, *tool usage*, *tokens* LLM), o que pode dificultar *debugs* no futuro.
3. Devido a rapidez do desenvolvimento, há a ausência de procedimentos de anonimização para PII.
4. Sem telemetria nas plataformas GA4/VTEX do *widget* para fazer o acompanhamento de métricas importantes (exibição, cliques, conversões assistidas).

Essa arquitetura privilegia simplicidade e tempo de implantação, mantendo *RAG-first* para reduzir riscos de conteúdo e integrações oficiais (VTEX/Correios) como fonte de verdade. À medida que os resultados forem medidos, o objetivo seria evoluir para telemetria completa e anonimização no n8n e atualização automática da base vetorial, consolidando a solução para suportar crescimento de tráfego e escopo (WhatsApp, novas *tools* e *playbooks* de exceção).

Pipeline das operações:

A) Recomendação de produto:

1. **RAG** (documentos + ficha de produto).
2. **VTEX API** (query do produto no site)
3. **Responder** de forma objetiva, adicionando a foto do produto e o link de compra

Pseudo-código:

```
q_norm = normalizar(consulta); q_enc = url_encode(q_norm)
atts = RAG_extrair_atributos(q_norm)      // categoria, uso, cor, tamanho
cand = vtex.search(q_enc)                  // candidatos "live"
cand_f = filtrar_por_categoria(cand, atts.categoria)
ranked = cohere.rerank(contexto=atts, candidatos=cand_f, topN=3)
ok = checar_estoque(ranked)
res = montar_resposta(ok, mostrar_preco=false) // link + imagem + variações
log(q_norm, atts.categoria, ids(ok))
```

B) FAQ / Políticas / Produto (conteúdo) - sequência adotada

1. **RAG** (documentos + ficha de produto).
2. **Citar fonte** (ID/slug interno e seção; incluir versão quando disponível).
3. **Responder** de forma objetiva; se a dúvida exigir **ação**, o agente **roteia** para a *tool* correspondente.
4. **Logging**: slug/versão do documento (sem PII).

Pseudo-código (conteúdo):

```
ctx = RAG_buscar(pergunta)
ref = extrair_slug_secao(ctx)
res = gerar_resposta(ctx, com_citacao=ref)
if exige_acao(res): roteador.encaminhar(tool_alvo)
log(slug=ref.slug, versao=ref.v)
```

C) WISMO (onde está meu pedido)

1. **Solicitar ID do pedido**
2. **Procurar Pedido (VTEX OMS)** > status, promessas e tracking (fonte canônica).
3. **Correios SRO** somente se o OMS não tiver evento atualizado (fallback).
4. **Resposta:** status atual, **próximo evento previsto e janela estimada**.
5. **Logging:** order_id, transportadora, atraso estimado, se houve oferta proativa. (E-mail para CS apenas se alguma regra de exceção disparar.)

Pseudo-código (WISMO):

```

id = obter_IDPedido(max_tentativas)
if !id: return responder("Não consegui localizar seu pedido; tente novamente com o
ID.")

oms = vtex.oms.get_order(id)
sro = (atualizado(oms)) ? oms.tracking : correios.sro(oms.trackingNumber)
if atraso(sro) > SLA_X: oferecer_notificacao_proativa()
responder(formatar_status(sro))
log(order_id=id, atraso=calc_atraso(sro), proativo=oferecido?)

```

D) Reversa (troca/devolução)

1. **Checar política/eligibilidade:** exige ID do pedido; confirmar entrega ≤ 7 dias para logística reversa gratuita.
 - 8-30 dias: informar envio por conta do cliente e endereço fixo.
2. **Se elegível (≤ 7 dias):** coletar TODOS os campos exigidos (nome, endereço, nº, bairro, cidade, UF, CEP, DDD, telefone, e-mail, qtdItems, descObjeto, NF, numeroPedido, valorDeclarado, descricaoInstrucao).
3. **Validar:** CEP (8 dígitos), DDD (2), telefone (8-9), valorDeclarado (decimal).
4. **Confirmar resumo** (cidade/UF, descObjeto, qtdItems, motivo, valorDeclarado).
5. **Solicitação Correios** (emitir autorização).
6. **Registrar deadline de 10 dias** (validade),
7. **Instruções de postagem claras** (o que levar, embalagem, prazo).
8. **Agendar lembretes T-48h e T-24h** (mensagens proativas).

9. **Acompanhar** (acompanhaPedido/tracking reverso)
10. **Escalonamento humano**: quando houver avaria/dano (análise de fotos).
11. **Notificações**: sempre e-mail ao CS e ao cliente.
12. **Logging**: **sem PII** (armazenar somente placeholders + IDs).

Pseudo-código (reversa):

```

id = exigir_IDPedido()
elig = checar_eligibilidade(id)
if dias_desde_entrega(id) <= 7:
    dados = coletar_campos_obrigatorios()
    if !validar(dados): return pedir_correcoes()
    confirmar_resumo(dados)
    auth = correios.reversa.solicitar(dados)
    registrar_deadline(auth, +10d)
    instruir_postagem(auth)
    agendar_lembretes(auth, T-48h, T-24h)
    acompanhar(auth)
    fechar_quando_postado(auth)
else if dias <= 30:
    informar_endereco_envio proprio()
else:
    informar_fora_de_prazo()
log(id, auth?.id, sem_PII=true)

```

Pipeline off-line (ingestão/atualização de conhecimento)

- **Produtos e documentos.** A coleção vetorial (Supabase/pgvector) agrega produtos/SKUs e documentos (suporte, trocasDevolucoes, FAQ, lojas, politica de privacidade).
- **Atualização:** incremental manual por CSV, mensal, executada pelo Gerente de TI.

- **Pré-processamento:** conforme o AS-IS, sem validador de amostra nem versionamento de índice; *embeddings* OpenAI text-embedding-3-small e re-ranking Cohere *top-3* com filtro por categoria.
- **Observação:** (*backlog*) incluir tabela de medidas e composição como metadados vetoriais para melhorar atributo/ajuste nas recomendações.

Observabilidade, logging e qualidade

- **Logging atual (mínimo):** uuid, sessionNumber, mensagem.
- **Acompanhamento de operação:** por n8n (regras, falhas de validação, tempos de ciclo, taxa de fallback).
- **Revisão humana: quinzenal,** conduzida por Gerente de TI e autor
- **Ajustes de prompt/políticas:** aprovados por autor + Gerente de TI + Gerente de Marketing.

Resiliência, escalonamento e comunicação

- **Erro em WISMO (pedido não localizado):** o agente informa que não conseguiu obter e solicita nova tentativa com ID.

Escalonamento humano: obrigatório em reversa com dano/avaria; demais exceções vão por e-mail ao CS.

- **Notificações transacionais:**
 - **Reversa:** sempre notifica cliente e CS por e-mail.
 - **WISMO:** notificação proativa ao cliente quando a regra de atraso > Xh aciona; cópia ao CS se houver exceção.
- **Canais:** por enquanto apenas **e-mail**. Planos futuros de expansão para o WhatsApp.

Ferramentas

No contexto deste trabalho, foi utilizado o *framework* da LangChain para construção do agente. Além disso, se destacam as seguintes *tools* implementadas no *bot*:

- “Procurar Pedido”: consulta ao OMS da VTEX para localizar pedidos e, quando aplicável, integração com a API dos Correios para rastreamento da encomenda.
- “Procurar Produto no Site”: utilização da API de busca da VTEX para localizar itens e sustentar recomendações baseadas em resultados, atributos e disponibilidade.
- “Solicitação Correios”: integração com a API de logística reversa dos Correios para abertura e acompanhamento de solicitações de devolução/troca.

Como infraestrutura de “conhecimento”, se adotou o Supabase: (i) como base vetorial para produtos (atributos, descrições) e documentos de normas (políticas, FAQs); (ii) com re-ranking aplicado ao *document retrieval*, otimizando precisão do contexto enviado ao modelo; e (iii) como repositório operacional de memória do *bot* e *log* de mensagens, favorecendo governança, diagnóstico e melhoria contínua.

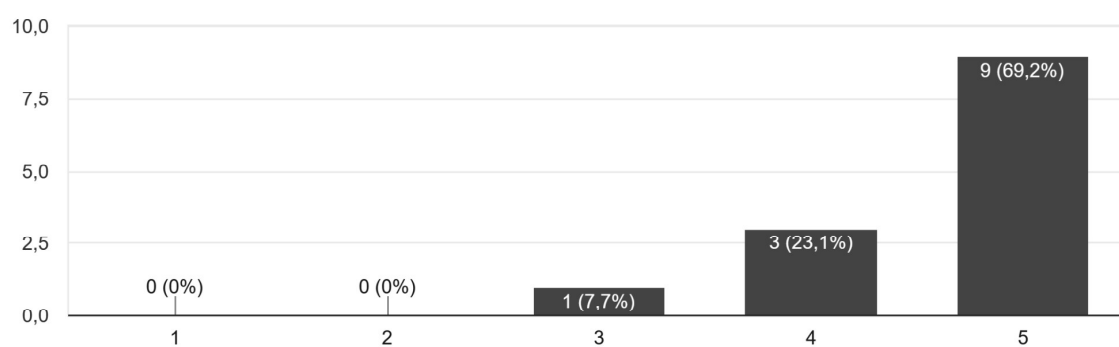
Apêndice B: Resultado do Survey no Google Forms

Este apêndice apresenta os resultados completos dos surveys aplicados aos usuários durante a fase de avaliação do Agente de IA da Brofitwear. Os gráficos foram extraídos diretamente do Google Forms, preservando escala, quantidades e percentuais.

Figura A.1 - Distribuição das respostas para “O agente foi fácil de usar”

O agente foi fácil de usar.

13 respostas

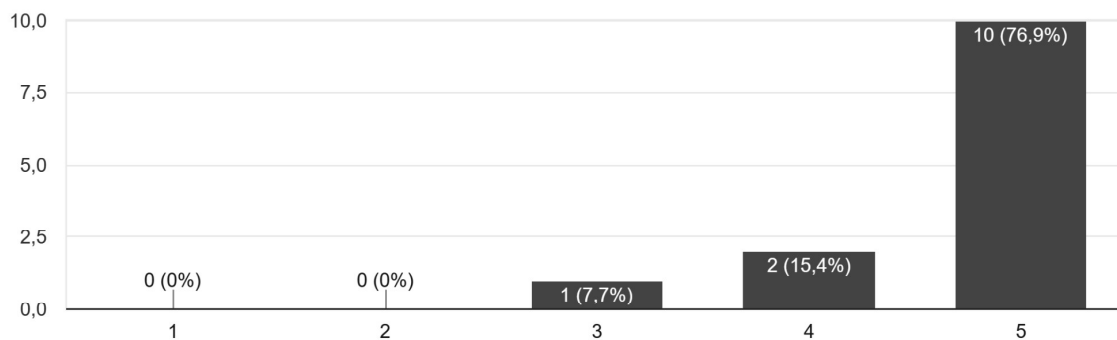


Fonte: Autoria Própria

Figura A.2 - Distribuição das respostas para “As respostas eram claras e compreensíveis.”

As respostas eram claras e compreensíveis.

13 respostas

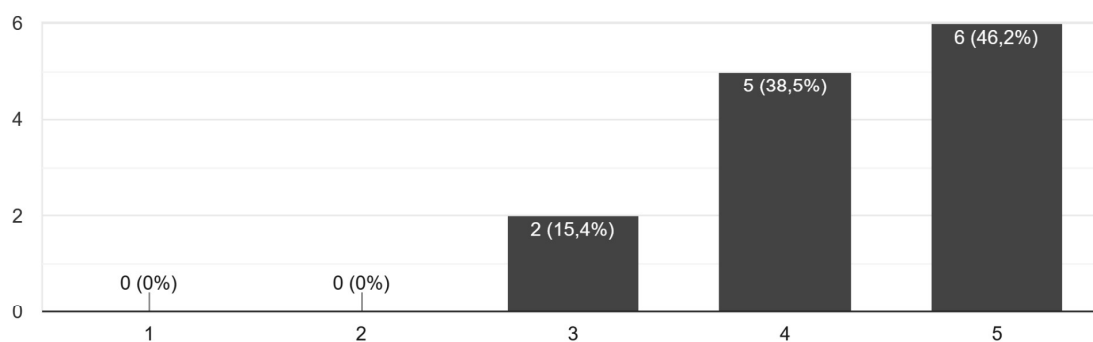


Fonte: Autoria Própria

Figura A.3 - Distribuição das respostas para “O fluxo de conversa foi natural e coerente.”

O fluxo de conversa foi natural e coerente.

13 respostas

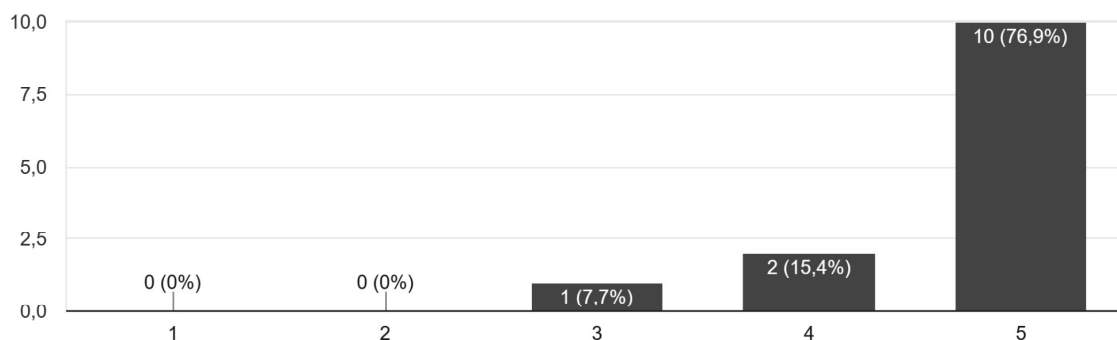


Fonte: Autoria Própria

Figura A.4 - Distribuição das respostas para “A linguagem usada pelo agente foi apropriada ao contexto da marca.”

A linguagem usada pelo agente foi apropriada ao contexto da marca.

13 respostas

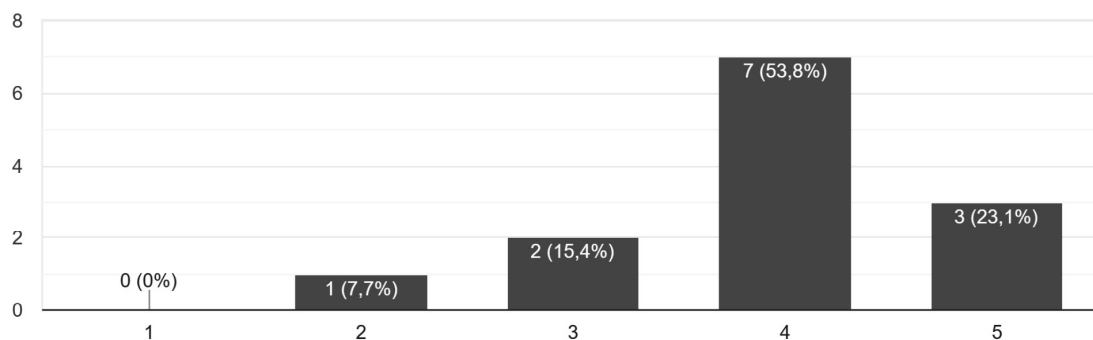


Fonte: Autoria Própria

Figura A.5 - Distribuição das respostas para “O tempo de resposta foi adequado. ”

O tempo de resposta foi adequado.

13 respostas

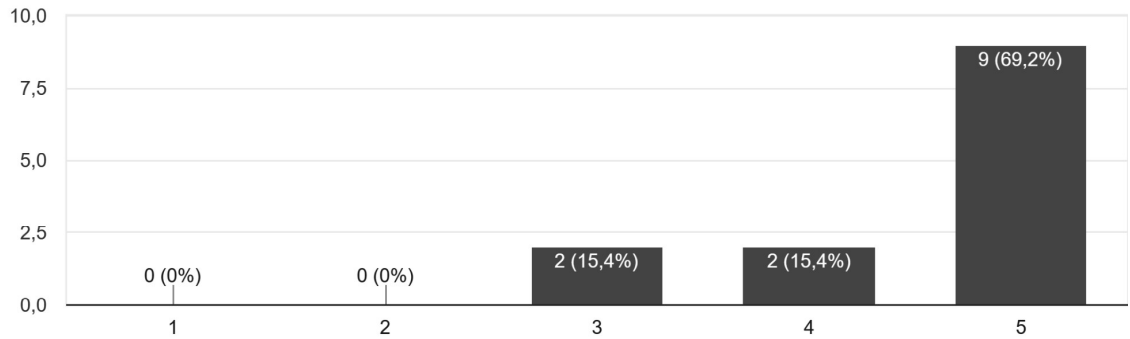


Fonte: Autoria Própria

Figura A.6 - Distribuição das respostas para “Consegui resolver minha necessidade sem ajuda humana. ”

Conseguir resolver minha necessidade sem ajuda humana.

13 respostas

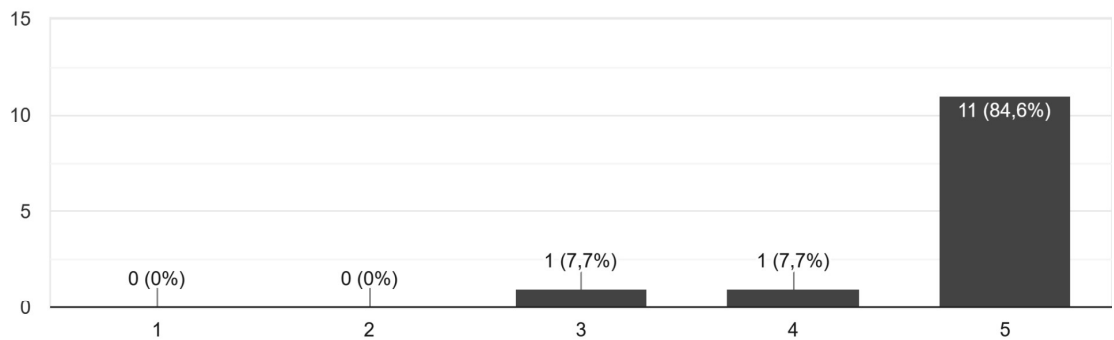


Fonte: Autoria Própria

Figura A.7 - Distribuição das respostas para “ O agente poupou meu tempo em comparação com o atendimento humano.”

O agente poupou meu tempo em comparação com o atendimento humano.

13 respostas

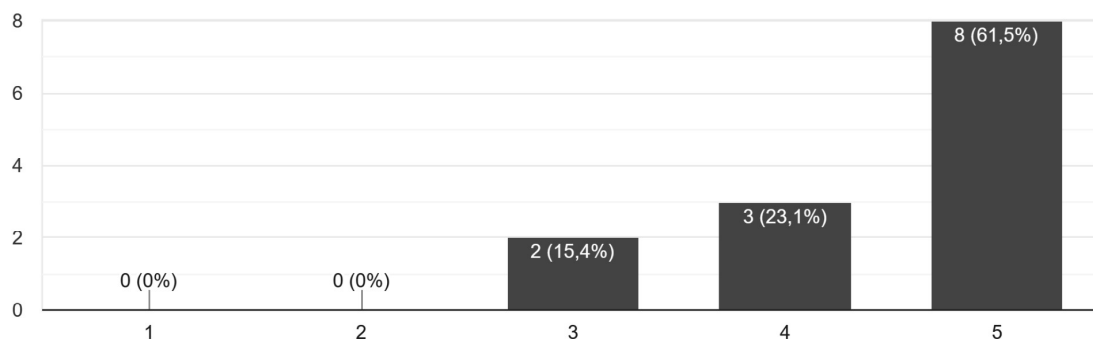


Fonte: Autoria Própria

Figura A.8 - Distribuição das respostas para “O agente entendeu corretamente o que eu queria. ”

O agente entendeu corretamente o que eu queria.

13 respostas

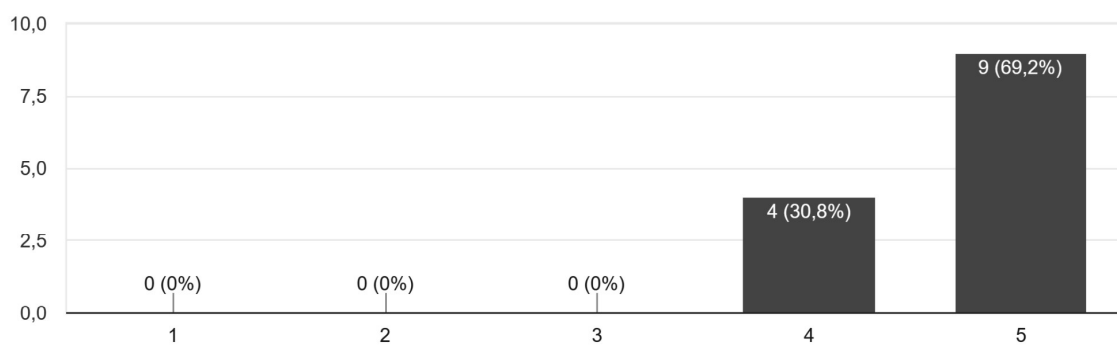


Fonte: Autoria Própria

Figura A.9 - Distribuição das respostas para “ As informações fornecidas foram corretas e relevantes. ”

As informações fornecidas foram corretas e relevantes.

13 respostas

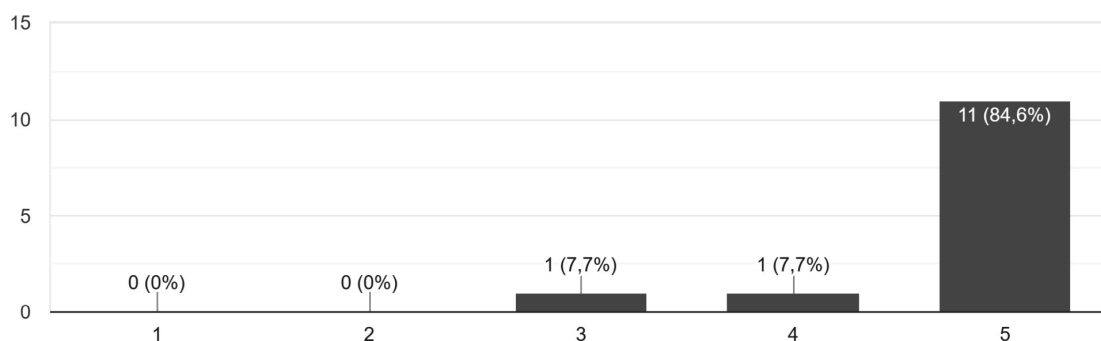


Fonte: Autoria Própria

Figura A.10 - Distribuição das respostas para “ As respostas foram transparentes sobre o que o agente podia ou não fazer.”

As respostas foram transparentes sobre o que o agente podia ou não fazer.

13 respostas

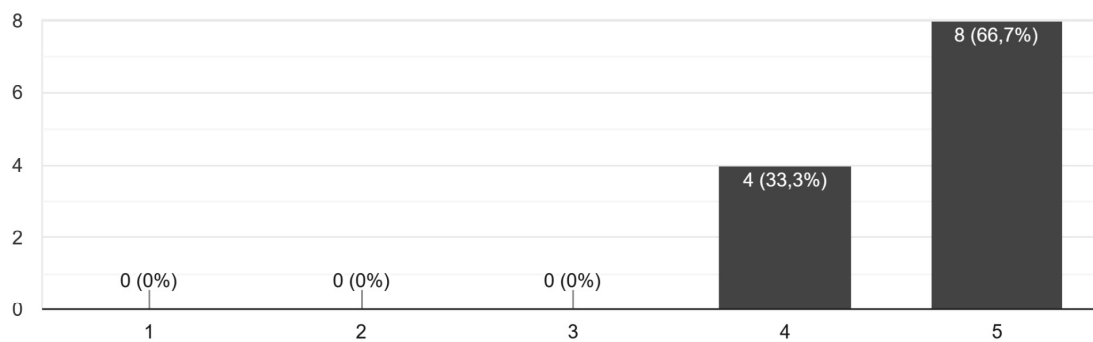


Fonte: Autoria Própria

Figura A.11 - Distribuição das respostas para “A integração com os sistemas (pedido, política, estoque) funcionou bem. ”

A integração com os sistemas (pedido, política, estoque) funcionou bem.

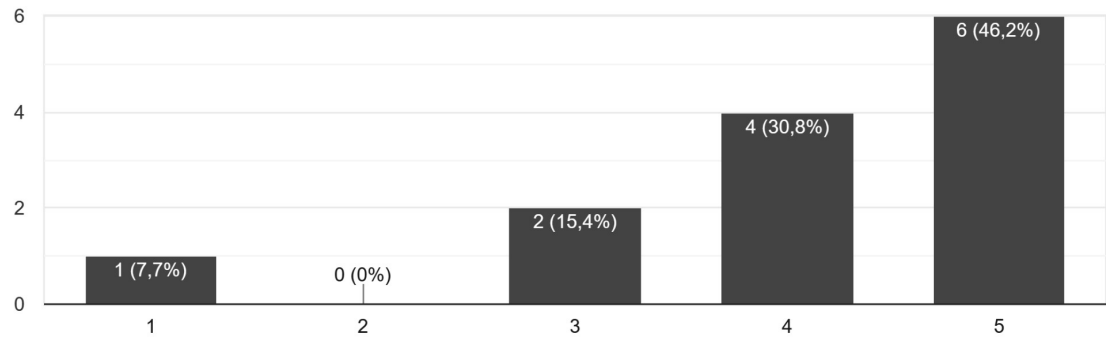
12 respostas



Fonte: Autoria Própria

Figura A.12 - Distribuição das respostas para “ Não encontrei dificuldades técnicas (erros, lentidão, travamentos)..”

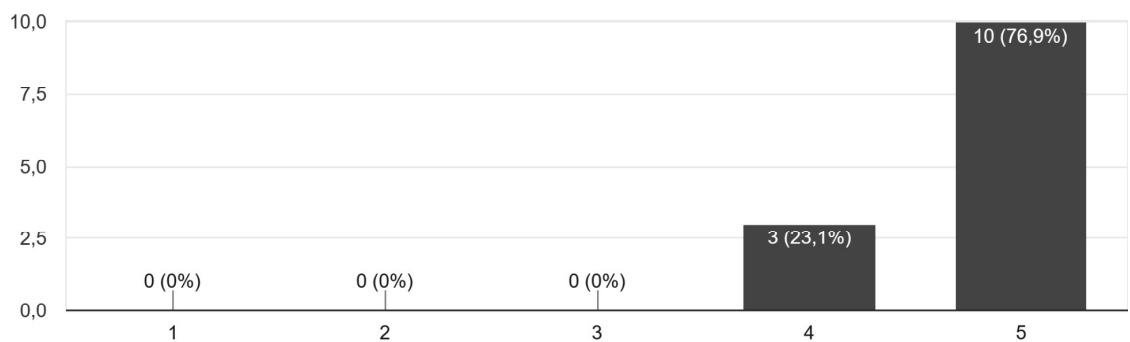
Não encontrei dificuldades técnicas (erros, lentidão, travamentos).
13 respostas



Fonte: Autoria Própria

Figura A.13 - Distribuição das respostas para “Eu usaria novamente o agente em futuras interações. ”

Eu usaria novamente o agente em futuras interações.
13 respostas

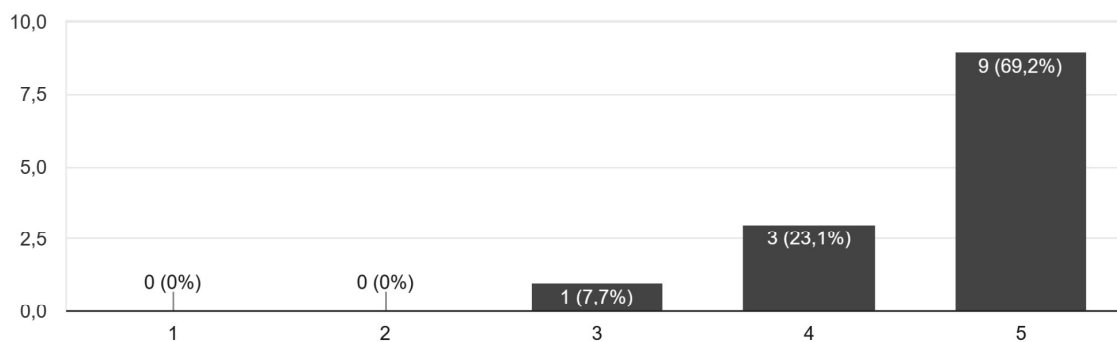


Fonte: Autoria Própria

Figura A.14 - Distribuição das respostas para “Recomendaria o uso desse agente para clientes reais. ”

Recomendaria o uso desse agente para clientes reais.

13 respostas

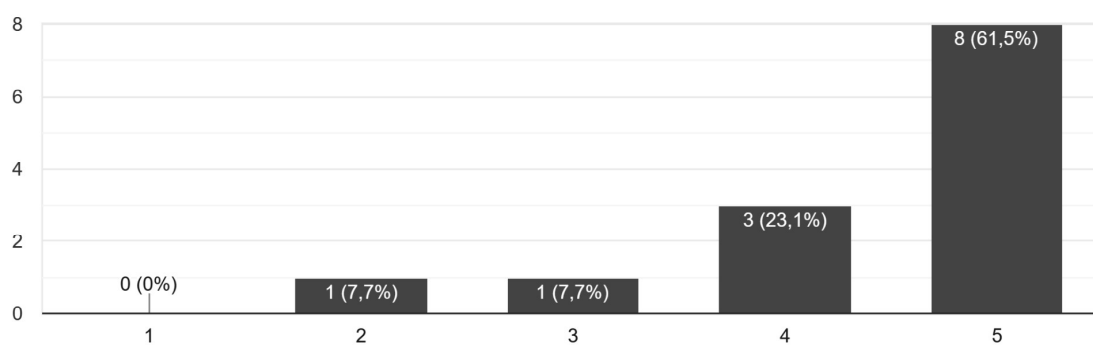


Fonte: Autoria Própria

Figura A.15 - Distribuição das respostas para “ Estou satisfeito com a experiência geral. ”

Estou satisfeito com a experiência geral.

13 respostas



Fonte: Autoria Própria