Impulse.

COLEÇÃO IMPULSE

Nº 01

ATENDIMENTO E RELACIONAMENTO

Atendente de IA no WhatsApp

que sabe chamar uma pessoa

Monte um agente na API oficial com n8n que responde, transfere para a equipe e respeita a janela de 24 horas.

O CAMINHO EM 3 PASSOS

1 bot → 2 fila → 3 pessoa

✓ Guia completo ✓ Planilha pronta ✓ Linguagem simples

PDF · EPUB (previsto)

1ª edição (em revisão)

Sumário

Esta é a **amostra gratuita**: contém os capítulos 1 e 6 completos. O livro inteiro tem dez capítulos e os anexos com código e testes.

> Um guia de implementação para colocar um atendente de IA no WhatsApp sem perder o controle da conversa: API oficial, n8n, base de conhecimento, transferência para humano e testes. Cada capítulo termina dizendo o que foi conferido, o que foi testado e o que ainda não foi validado.

01 **O problema e quando o agente nativo do WhatsApp Business basta** — nesta amostra

02 Arquitetura: as peças e o caminho de uma mensagem

03 Preparação: contas, número, permissões, custos e ambiente de teste

04 Base de conhecimento

05 O primeiro fluxo: receber, entender, responder e registrar

06 **Quando o bot deve chamar uma pessoa** — nesta amostra

07 Mensagens fora da janela: modelos aprovados

08 Testes, critérios de aceitação e entrada em produção

09 Privacidade, LGPD e controle de acesso

10 Indicadores, revisão semanal e manutenção

A Anexos: código, testes, SQL, modelos de mensagem e checklist

Versão 0.2: todos os capítulos escritos. Antes de "pronto para publicação", falta executar os seis cenários ponta a ponta do capítulo 8 num ambiente de teste real e a revisão técnica cruzada. Exemplos, empresas e números são fictícios.

CAPÍTULO 1

O problema e quando o agente nativo do WhatsApp Business basta

> Ao fim deste capítulo, você saberá dizer se precisa do que este livro ensina a montar, ou se uma solução mais simples resolve o seu caso. Muitas empresas não precisam, e é melhor descobrir isso antes de gastar uma semana.

1.1 O que dá errado no atendimento pelo WhatsApp

O WhatsApp virou o balcão de muitas empresas brasileiras. A mesma conversa mistura pergunta de preço, pedido de orçamento, reclamação e cliente antigo pedindo segunda via. Quando o volume cresce, aparecem três problemas conhecidos:

- **Demora.** Mensagens fora do horário ficam sem resposta até a manhã seguinte; no horário, a fila cresce nos picos.
- **Respostas diferentes para a mesma pergunta.** Cada atendente responde do seu jeito, com a informação que lembra.
- **Ninguém sabe o que acontece.** Sem registro, não dá para dizer quantas conversas chegaram, quais viraram venda e o que os clientes mais perguntam.

Um atendente de IA resolve parte disso: responde de imediato as perguntas que já têm resposta oficial e organiza o restante para a equipe. O que ele não pode fazer é responder com segurança o que não sabe, nem atropelar um atendente humano. Este livro trata justamente desses dois limites.

1.2 Três caminhos possíveis

Caminho	O que é	Serve quando
Recursos do aplicativo	Mensagem de saudação, mensagem de ausência, respostas rápidas, catálogo e listas do WhatsApp Business, sem IA.	O volume é pequeno e uma pessoa responde ao longo do dia.
Agente nativo da Meta	Agente de IA dentro do próprio aplicativo WhatsApp Business, lançado no Brasil em fevereiro de 2026. Responde dúvidas e pode passar a conversa para você.	Uma pessoa atende pelo celular, sem necessidade de consultar outros sistemas.
Montagem deste livro	WhatsApp Cloud API + n8n + banco de dados + modelo de IA, com regras de transferência sob seu controle.	Várias pessoas atendem, o agente precisa consultar pedidos, agenda ou CRM, ou você precisa de registros e indicadores próprios.

Fontes: página oficial do agente de IA do WhatsApp Business (whatsappbusiness.com/pt-br/products/business-app-ai-agent), consultada em 15/10/2026; Forbes Brasil, 24/02/2026. O agente nativo não foi testado pela Impulse e a página não informava preço na data da consulta.

1.3 Um teste de cinco perguntas

Responda com sim ou não:

1. Mais de uma pessoa responde clientes no mesmo número?
2. O cliente pergunta coisas que só um sistema sabe (status do pedido, horário livre na agenda, saldo, segunda via)?
3. Você precisa de regras próprias para decidir quando uma pessoa assume (por tema, por horário, por valor)?
4. Você quer guardar as conversas no seu banco para medir e melhorar (capítulo 10)?
5. Há alguém na empresa, ou um parceiro, capaz de manter um fluxo no n8n?

Nenhum "sim" nas quatro primeiras: comece pelos recursos do aplicativo ou pelo agente nativo. **Um ou mais "sim" nas quatro primeiras e "sim" na quinta:** este livro é para você. **"Sim" nas primeiras e "não" na quinta:** a montagem é possível, mas planeje quem vai cuidar dela depois; um atendente sem manutenção piora com o tempo.

1.4 O que você terá ao final do livro

- Um número de WhatsApp na API oficial, com um atendente que responde só com base numa base de conhecimento revisada.
- Uma regra de transferência testada, que chama uma pessoa nos casos certos e fica em silêncio enquanto ela atende.
- Modelos de mensagem aprovados para falar com clientes fora da janela de 24 horas.
- Um roteiro de testes de aceitação, um checklist de entrada em produção e consultas prontas para os indicadores semanais.

NÃO USE BIBLIOTECAS NÃO OFICIAIS

Há ferramentas que conectam robôs ao WhatsApp imitando o aplicativo (lendo o QR code do WhatsApp Web). Elas não são a interface oficial da Meta e expõem o número a bloqueio. Este livro usa apenas a WhatsApp Cloud API.

ESTADO DE VALIDAÇÃO DESTE CAPÍTULO

- **Conferido (15/10/2026):** existência e descrição do agente nativo na página oficial da Meta.
- **Não validado:** funcionamento e custo do agente nativo; o teste de cinco perguntas é um critério prático da Impulse, não uma regra da Meta.

CAPÍTULO 6

Quando o bot deve chamar uma pessoa

> Ao fim deste capítulo, o seu agente vai saber quando parar de responder, avisar a equipe, ficar em silêncio enquanto um atendente conversa e voltar sozinho quando o atendente sair, sem desrespeitar a janela de 24 horas do WhatsApp.

6.1 O problema que este capítulo resolve

Um agente que responde tudo cria dois riscos. O primeiro é responder com segurança algo que ele não sabe, como prazo de um pedido específico ou condição de um reembolso. O segundo aparece quando um atendente entra na conversa: se o bot continua respondendo, o cliente recebe duas vozes ao mesmo tempo e perde a confiança nas duas.

A solução não está no texto do *prompt* (a instrução dada ao modelo de IA). Está em uma regra explícita, fora do modelo, que diz em que **estado** cada conversa se encontra e quem pode falar nele. O modelo continua escrevendo as respostas, mas não decide sozinho se deve falar.

6.2 Os quatro estados de uma conversa

Cada conversa fica sempre em exatamente um destes estados, gravado no banco de dados:

Estado	Quem fala	Sai dele quando
`bot`	A IA responde ao cliente.	O cliente pede uma pessoa, toca em tema sensível ou a IA não tem confiança na resposta.
`aguardando_humano`	Ninguém responde a cada mensagem. A equipe foi avisada.	Um atendente assume. Se ninguém assumir em 15 minutos, o cliente recebe um único aviso.
`humano`	Só o atendente. O bot apenas notifica a equipe de novas mensagens.	O atendente devolve a conversa ou passa 20 minutos sem responder.
`encerrada`	Ninguém.	O cliente volta a escrever, e a conversa recomeça no bot.

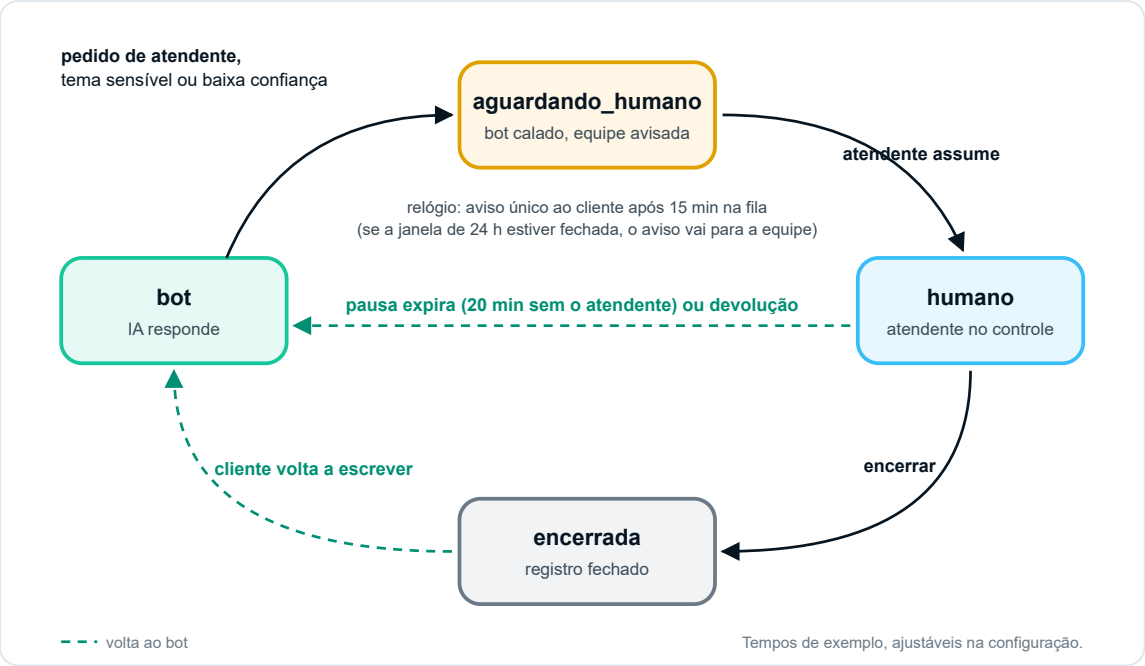


Figura 6.1 — Estados e transições. Os tempos de 15 e 20 minutos são valores iniciais de exemplo; ajuste-os à rotina da sua equipe.

6.3 Regras de transferência

Comece com três gatilhos simples e mensuráveis. Regras simples são fáceis de testar e de explicar à equipe.

1. **Pedido explícito.** O cliente escreve palavras como "atendente", "humano" ou "falar com uma pessoa". Nunca obrigue alguém a convencer o bot de que precisa de gente.
2. **Tema sensível.** Reclamação, reembolso, cancelamento, menção a órgãos de defesa do consumidor. Ajuste a lista ao seu negócio.
3. **Baixa confiança.** No capítulo 5, o fluxo pede ao modelo uma nota de 0 a 1 indicando se a resposta está apoiada na base de conhecimento. Abaixo de 0,6, a conversa vai para uma pessoa. A nota dada pelo próprio modelo é um sinal fraco; o capítulo 8 mostra como conferir esse limite com conversas reais.

LIMITE CONHECIDO DA REGRA POR PALAVRA

A busca por trecho de texto encontra "pessoa" também dentro de "pessoal". Isso gera transferências desnecessárias, o que é um erro mais barato do que deixar de transferir. O exercício 6.2 mostra como trocar a busca por palavra inteira e medir a diferença com as conversas da sua própria base.

6.4 A janela de 24 horas

Na API oficial do WhatsApp (WhatsApp Cloud API), mensagens livres, que não sejam um modelo aprovado (*template*), só podem ser enviadas com a **janela de atendimento** aberta. A janela se abre

quando o cliente escreve e dura 24 horas a partir da última mensagem dele. Fechada a janela, a empresa só pode iniciar contato com um modelo aprovado pela Meta.

Isso afeta a transferência de duas formas: um atendente que demora mais de 24 horas não pode responder com texto livre, e o aviso automático "você continua na fila" não pode sair depois que a janela fechou. A máquina de estados trata os dois casos.

Fonte: documentação de preços e janelas da WhatsApp Business Platform, developers.facebook.com/docs/whatsapp/pricing, consultada em 15/10/2026. Reconferida em 16/10/2026: a página oficial não lista mudanças para o Brasil nem para mensagens de atendimento em 2026–2027 (detalhes no capítulo 3).

6.5 A regra em código

A regra fica em uma função pura: recebe a conversa e um evento e devolve a conversa atualizada e uma lista de ações. Ela não envia mensagens nem consulta o relógio do servidor. Por isso pode ser testada sem WhatsApp, sem n8n e sem internet. O arquivo completo, `transferencia.mjs`, acompanha o livro. Veja o trecho central:

```
case 'mensagem_cliente': {
  c.ultimaMsgCliente = agora;                // reabre a janela de 24 h
  if (c.estado === ESTADOS.BOT) {
    const motivo = motivoParaTransferir(evento.texto, evento.confianca, config);
    if (motivo) {
      c.estado = ESTADOS.FILA;
      c.entrouNaFila = agora;
      acoes.push({ tipo: 'avisar_equipe', motivo });
      acoes.push({ tipo: 'responder', modelo: dentroDoExpediente(agora, config)
        ? 'transferindo' : 'transferindo_fora_expediente' });
    } else {
      acoes.push({ tipo: 'responder_com_ia' });
    }
  } else if (c.estado === ESTADOS.HUMANO) {
    acoes.push({ tipo: 'notificar_atendente' });   // o bot fica calado
  }
  break;                                     // na fila: silêncio; o aviso sai pelo relógio
}
```

Dois detalhes evitam erros comuns. O expediente é calculado com o fuso informado na configuração, e não com o fuso da máquina onde o n8n roda, que costuma estar em UTC. Um evento desconhecido gera erro em vez de ser ignorado: falha silenciosa em atendimento é cliente sem resposta.

6.6 Como o n8n usa a função

No n8n, a função roda dentro de um nó *Code*. O fluxo de mensagens recebidas fica assim:

1. **WhatsApp Trigger** recebe a mensagem (evento *Messages*).
2. **Postgres** busca a conversa pelo número do cliente, ou cria uma nova.
3. **Code** chama `transicao(conversa, evento, Date.now())`.

4. **Postgres** grava o novo estado.
5. **Switch** separa as ações: `responder_com_ia` segue para o agente do capítulo 5; `avisar_equipe` envia a notificação ao grupo da equipe; `responder` envia o modelo de mensagem indicado.

Um segundo fluxo, com **Schedule Trigger** a cada minuto, envia o evento `relogio` para as conversas nos estados `aguardando_humano` e `humano`. É ele que expira a pausa e dispara o aviso de fila.

> **UM WEBHOOK POR APLICATIVO**
>
> A documentação do n8n alerta que o WhatsApp registra um único webhook por aplicativo da Meta. Ao ativar o fluxo de teste, o de produção deixa de receber mensagens. Use um aplicativo separado para testes ou desative o fluxo de produção durante o teste.

6.7 Testes e critérios de aceitação

A transferência está pronta para produção quando todos os casos abaixo passam. Os onze testes automáticos que acompanham o livro cobrem cada linha:

Situação	Resultado esperado
Pergunta comum, confiança 0,9	IA responde; estado continua `bot`
"Quero falar com um atendente"	Fila, equipe avisada, cliente informado
Confiança 0,3	Fila, motivo `baixa_confianca`
Pedido de reembolso no sábado	Fila, resposta com horário da equipe
Cliente insiste na fila	Bot calado; um único aviso após 15 min
Cliente escreve com atendente no controle	Bot calado; atendente notificado
Atendente some por 20 min	Conversa volta ao bot e o evento é registrado
Atendente responde após 25 h	Sistema exige modelo aprovado
Janela fechada durante a fila	Aviso vai para a equipe, não para o cliente
Conversa encerrada recebe "oi"	Recomeça no bot
Evento desconhecido	Erro explícito

```
# na pasta dos materiais do livro
node --test codigo/transferencia.test.mjs
# resultado esperado: pass 11, fail 0
```

6.8 Falhas comuns e diagnóstico

Sintoma	Causa provável	O que verificar
Bot responde junto com o atendente	Mensagem do atendente não gera o evento `humano_respondeu`	Confirme que cada resposta do atendente chega ao n8n como evento. Se a equipe responde pelo aplicativo do celular no mesmo número, com a coexistência, a Meta envia cada uma delas à API como webhook `smb_message_echoes` (capítulo 3); trate esse evento como `humano_respondeu`. Sem coexistência, responda por uma caixa de atendimento integrada.
Avisos de fila fora de hora	Fuso horário	Valor de `fusoMin` e hora do servidor do n8n
Mesma mensagem tratada duas vezes	Reenvio do webhook	Grave o identificador de cada mensagem recebida e ignore repetições
Fluxo de produção parou após um teste	Webhook único por aplicativo	Qual URL está registrada no painel da Meta
Erro ao responder um cliente antigo	Janela de 24 h fechada	Use um modelo aprovado (capítulo 7)

6.9 O que continua sendo decisão humana

A máquina decide quem fala, mas não decide reembolsos, exceções de prazo, descontos ou respostas a reclamações formais. Essas decisões ficam com a equipe, e o agente não deve prometê-las. Registre o motivo de cada transferência: no capítulo 10, esse registro mostra quais perguntas a base de conhecimento ainda não cobre.

Exercícios

1. **6.1** Liste os cinco temas que, no seu negócio, sempre exigem uma pessoa. Adicione-os à configuração e escreva um teste para cada um.
2. **6.2** Troque a busca por trecho por uma busca por palavra inteira. Rode as duas versões sobre 50 mensagens reais (sem nomes nem telefones) e compare as transferências.
3. **6.3** Escolha os tempos de fila e de pausa a partir do tempo médio de resposta da sua equipe, não dos valores deste capítulo.

ESTADO DE VALIDAÇÃO DESTE CAPÍTULO

- **Testado:** `transferencia.mjs` com os 11 testes automáticos, Node.js 24.19, em 15/10/2026. Todos passaram.
- **Conferido na documentação oficial (15/10/2026):** regra da janela de atendimento (Meta) e limite de um webhook por aplicativo (n8n).
- **Conferido na documentação oficial (16/10/2026):** mensagens enviadas pelo aplicativo em coexistência chegam à API (`smb_message_echoes`).
- **Ainda não validado:** o fluxo n8n da seção 6.6 importado e executado com a WhatsApp Cloud API e um número de teste; reenvio de webhook observado na prática.
- **Exemplos:** números, horários e mensagens são fictícios e servem só para demonstração.