

Guia de Agentes de IA

Agentes de IA

O Agente de IA é um atendente virtual que conversa com seus clientes pelo WhatsApp, 24 horas por dia, usando as informações e as regras do **seu** negócio. Ele qualifica o lead, responde dúvidas com base nos seus documentos, envia botões e listas, aplica etiquetas, faz follow-up de quem parou de responder e transfere para a sua equipe na hora certa — tudo dentro de uma fila de atendimento que você escolhe.

 [Captura de tela aqui — tela Agentes de IA com um agente cadastrado]

Antes de começar

Você vai precisar de:

- **Uma conexão de WhatsApp ativa e uma fila de atendimento** dedicada ao agente (menu Filas de Atendimento). Todo cliente que cair nessa fila será atendido pela IA.
- **Um modelo de IA:** informe sua própria chave de API (OpenAI, Google ou Anthropic) — ou deixe o campo de chave em branco para usar a **IA gerenciada da plataforma**, que consome créditos do seu plano sem precisar configurar nada.

Criando o agente

Acesse **IA & Automação → Agentes de IA → Adicionar**.

 [Captura de tela aqui — modal de criação do agente]

- **Nome e fila:** o nome é interno; a fila define quais conversas o agente atende.
- **Provedor, modelo e chave de API:** escolha o provedor e digite o ID do modelo (ex.: `gemini-2.5-flash`, `gpt-4o-mini`). Sem chave = modo gerenciado (créditos).
- **Instruções (prompt):** a alma do agente. É aqui que você escreve quem ele é, como fala, o que pergunta, quando transfere e todas as regras do seu negócio. Veja as boas práticas no fim desta página.

Bases de Conhecimento

Documentos que o agente **consulta** para responder: catálogos, tabelas, FAQ, descrições de produtos ou serviços. Crie em **IA & Automação → Bases de Conhecimento**, envie os arquivos e depois marque a base no cadastro do agente.

- O conteúdo é dividido em trechos automaticamente. Para controlar a divisão, separe blocos com uma linha contendo apenas `---` — cada bloco vira um trecho inteiro.
- **Regra de ouro:** o que é *conteúdo consultável* (tabelas, descrições, listas) vai para a base; o que é *regra de comportamento* ("nunca faça X", "sempre pergunte Y") vai para o prompt.

 [Captura de tela aqui — tela Bases de Conhecimento]

Ferramentas do Agente (consultas externas)

Para dados que mudam o tempo todo — estoque, preço em sistema externo, disponibilidade de agenda — cadastre uma **ferramenta de consulta** em **IA & Automação → Ferramentas do Agente**. O agente chama o seu endpoint (n8n, ERP, API própria) **antes** de responder e usa o dado exato em vez de estimar.

- A **descrição** da ferramenta é lida pela IA — é ela que decide quando chamar. Escreva como uma instrução: "*Consulta preço e disponibilidade de um produto pelo nome. Use SEMPRE antes de informar preço ao cliente.*"
- Uma **calculadora** já vem embutida em todo agente: contas de soma, multiplicação e porcentagem são feitas com precisão, sem você configurar nada.

 [Captura de tela aqui — modal de nova ferramenta]

Ações no CRM

Na aba de ferramentas do cadastro do agente, você escolhe o que ele pode fazer **além de conversar**:

- **Aplicar/remover etiquetas** — o agente só consegue aplicar etiquetas que estejam **citadas nas instruções dele**. Etiqueta que o prompt não menciona é recusada automaticamente (proteção contra marcação indevida).

- **Transferir para atendente / atribuir atendente disponível**
- **Encerrar o atendimento**
- **Agendar mensagem e criar tarefa** para a equipe
- **Mensagens interativas** — com esta opção ligada, o agente envia **botões e listas** sozinho sempre que oferece uma escolha fechada ao cliente (períodos, opções, confirmação de resumo). Botão engaja muito mais que pergunta aberta.
- **Nota interna** — registros visíveis só para a equipe.

☐☐ [Captura de tela aqui — aba de ferramentas do agente]

Transferência para a equipe

Três formas de o atendimento chegar a um humano:

1. **Palavras-chave:** defina frases como "falar com um atendente" — ao detectá-las na mensagem do cliente, a transferência é **imediata**, sem passar pela IA. A comparação é por palavra/frase inteira ("sim" não dispara em "assim").
2. **Decisão do agente:** quando o cliente confirma uma proposta ou pede um humano, o agente transfere sozinho — com proteções para nunca transferir com uma pergunta ainda sem resposta.
3. **Botão de confirmação:** instrua o agente (no prompt) a enviar um botão após o resumo, com o texto igual a uma das suas palavras-chave — o toque do cliente vira transferência instantânea.

“☐☐ **Dica:** na mensagem de despedida do agente, nunca prometa prazo de resposta do atendente ("já te responde!") — pode ser madrugada ou fim de semana. Prefira "ele te responde por aqui mesmo".

Revisão automática (agente revisor)

Com a revisão ligada, **toda resposta passa por um segundo modelo de IA antes de ser enviada**, que confere o texto contra as suas regras e contra o material consultado na base — horários, valores, itens repetidos, sequências. Se encontrar erro, a resposta é reescrita na hora. É a diferença entre um agente que "quase sempre acerta" e um que você confia de olhos fechados.

- Você pode usar um modelo diferente (mais barato ou mais rigoroso) só para a revisão.

- A revisão tem custo próprio de tokens — em troca, pega erro factual antes do cliente ver.

Escuta ativa (plantão após transferência)

Com a **escuta ativa** ligada, o agente não sai da conversa quando um atendente assume o ticket — fica de plantão, em silêncio. Se o cliente mandar mensagem e o atendente estiver ausente (ou não responder em 5 minutos estando online), o agente responde na hora, com todo o contexto. Qualquer mensagem do atendente — inclusive enviada pelo celular — devolve a vez a ele.

Ideal para clientes que respondem fora do horário de atendimento.

Follow-up automático

Se o cliente parar de responder, o agente envia **mensagens de resgate geradas pela IA** — retomando a última pergunta que ficou aberta, nunca um template genérico.

☐☐ [Captura de tela aqui — seção de follow-up no cadastro do agente]

- **Esperas entre tentativas:** escolha um **preset pelo objetivo do agente** — *SDR / Qualificação* (30m, 3h, 18h), *Vendas / Negociação* (1d, 3d, 7d) ou *Suporte* (2h, 1d) — com um clique, ou personalize livremente (ex.: `30m, 3h, 18h`). A quantidade de esperas define o número de tentativas, e o tom evolui sozinho: primeira leve, intermediárias empáticas, última é uma despedida elegante que deixa a porta aberta.

☐☐ [Captura de tela aqui — chips de preset de cadência]

- **Dias e horário de envio:** fora da janela, a tentativa espera a próxima abertura (nada de mensagem de madrugada). O fuso horário usado é o da sua empresa — configure em **Configurações → Atendimento → Fuso horário**.
- **Encerrar ao final:** quem não responde a nenhuma tentativa tem o ticket encerrado sozinho — separa curiosos de clientes reais sem ocupar a equipe.
- Se o cliente responder a qualquer momento, o follow-up pendente é cancelado na hora.
- Cliente que responder depois do encerramento **reabre o ticket na mesma fila**, com o agente e todo o histórico.



⚠️ Vários follow-ups em sequência aumentam o risco de denúncia no WhatsApp.
3 a 4 tentativas é a boa prática.

Imagens e áudios

- **Imagens:** quando o cliente envia um print ou foto (comprovante, tabela, documento, tela de outro sistema), o conteúdo é **lido automaticamente** e vira contexto da conversa — o agente enxerga o que está escrito na imagem e usa a informação, inclusive nos turnos seguintes. Você pode instruir no prompt o que fazer com tipos específicos de imagem.
- **Áudios:** são transcritos e entram na conversa como texto.

Toques finais

- **Presença "digitando...":** o cliente vê o agente digitando antes de cada mensagem, como uma conversa humana.
- **Mensagens divididas:** respostas longas chegam em blocos naturais, com pausas — não em um "textão".

Boas práticas de prompt

1. **Persona primeiro:** nome, tom de voz, do que a empresa vive.
2. **Regras críticas em destaque:** use marcadores como `[]` e palavras como NUNCA/SEMPRE nas regras inegociáveis.
3. **Scripts para momentos-chave:** primeira mensagem, coleta de dados, resumo, despedida. Diga quando o script é obrigatório.
4. **Etiquetas:** cite pelo nome exato as etiquetas que o agente deve usar e quando. Ele só consegue aplicar as citadas.
5. **O que é da base fica na base:** prompt define comportamento; base guarda conteúdo.
6. **Teste como cliente:** converse de verdade pelo WhatsApp antes de liberar. Ajuste o prompt a cada rodada — pequenas frases mudam muito o comportamento.

Custos (ordem de grandeza)

Uma conversa completa de qualificação (7 a 10 mensagens, com revisão ligada) custa em torno de **R\$ 0,50 a R\$ 1,00** em tokens com um modelo econômico como o `gemini-2.5-flash`. Leitura de imagem: centavos por foto. No modo gerenciado, cada interação consome 1 crédito do plano.

Revision #2

Created 2026-08-30 23:24:31 UTC by Maurilio Vilela

Updated 2026-09-01 00:08:02 UTC by Maurilio Vilela