AstraChat e Agente de voz com IA: como ele atende, agenda e transfere

Link vídeo: https://youtu.be/vr8u9EQUkeE

Link tutorial: https://chatwoot.astraonline.com.br/blog/agente-de-voz-ia-atendimento/

Por Raphael Popiolek · 28 de outubro de 2026 · Atualizado em 03 de novembro de 2026

Um **agente de voz com IA** é uma IA que atende a ligação e conversa com o cliente por voz em tempo real: no AstraChat, ele consulta a agenda, marca o horário e, se o cliente pedir, passa a ligação para alguém da equipe.

Uma clínica ou uma barbearia que recebe ligação pelo WhatsApp costuma ter o mesmo problema: quem atende o telefone é a mesma pessoa que está no balcão, e a ligação fica sem resposta.

Na live eu passo por todos os tipos de Flow do AstraChat, e o último é justamente o agente de voz. Neste artigo eu foco nele: como as peças se conectam, o que pesa na latência e no custo, onde ele funciona bem e onde ainda tem limite. Se você procura um painel separado, montado do zero numa VPS, esse é outro projeto e está explicado em IA que atende ligação no WhatsApp.

https://www.youtube.com/embed/vr8u9EQUkeE?rel=0

Onde o agente de voz entra nos Flows

O construtor de Flows do AstraChat tem quatro tipos. Vale entender a diferença, porque o agente de voz reaproveita quase tudo dos outros:

Tipo de Flow	Quando começa	Para que serve
Chatbot	quando o cliente manda mensagem	menus, triagem, agente de IA por texto
Sequência	quando você coloca o cliente nele	follow-up com esperas e verificação de resposta
Workflow	por gatilho: etiqueta, etapa do Kanban, webhook, horário, resposta a disparo	automações internas
Voz	quando o cliente liga	a IA atende e conversa por voz

Os tipos podem funcionar juntos na mesma caixa. Você pode ter um chatbot de texto e um agente de voz no mesmo número, ou um agente de voz diferente para cada caixa.

A arquitetura, peça por peça

O caminho de uma ligação é este:

1. O cliente liga para o número conectado pelo **AstraCalls**, a nossa API de ligações. A caixa precisa estar com as ligações ativadas nas conexões.
2. A chamada toca rapidamente para o atendente e a IA assume.
3. O **nó de voz** conversa usando um modelo em tempo real, da OpenAI ou do Google Gemini.
4. Quando precisa agir, o agente usa os **nós de ação** ligados a ele como ferramentas: consultar disponibilidade, agendar, cancelar, atribuir a um atendente, chamar um sistema externo por HTTP.
5. Ao encerrar, ele manda uma mensagem pelo WhatsApp e pode salvar a transcrição no ticket.

No nó de voz você configura a voz, o modelo, a primeira frase que a IA fala ao atender, a orientação (o prompt), a mensagem pós-ligação, uma base de conhecimento e a duração máxima da chamada em segundos.

Na live, gravada em setembro de 2026, o agente de voz funcionava só com o AstraCalls, e a conexão com as caixas de API oficial estava em desenvolvimento. Ele também estava focado em agentes de IA, e não em tocar um áudio fixo.

O teste da live: agendamento e transferência

Usei um flow de agendamento de clínica. A IA atendeu, perguntou o serviço (avaliação, consulta inicial ou retorno), quis saber se eu tinha preferência de profissional ou data, ofereceu os dois profissionais livres no horário, pediu meu nome completo e confirmou. O agendamento apareceu na hora no módulo de atendimentos.

Depois pedi para falar com um atendente. A IA avisou que ia transferir e a chamada voltou a tocar no AstraChat para alguém da equipe continuar. Se ninguém assumir, a IA encerra e manda a mensagem de fechamento pelo WhatsApp.

Um detalhe que travou o teste por texto, um pouco antes: todos os profissionais estavam marcados como indisponíveis, então a IA nunca encontrava horário. Configure serviços, profissionais e disponibilidade antes de culpar o prompt.

Latência: por que só modelos de voz em tempo real

Numa ligação, um atraso de poucos segundos já deixa a conversa estranha. Por isso o nó de voz usa só modelos que recebem e devolvem áudio em tempo real, e não o caminho texto, resposta, conversão para áudio. Hoje são dois: OpenAI e Google Gemini.

Na minha experiência, a voz da OpenAI flui melhor, e o Gemini às vezes atrasa ou dá uma travada. Por outro lado, a OpenAI sai mais cara. Teste os dois com o seu roteiro antes de decidir. Se você conhece outro modelo com voz em tempo real que funcione bem, me conte, porque a lista pode crescer.

Do lado do servidor, a parte de voz é leve. O peso fica no modelo, que roda na nuvem do provedor.

Custos: o que entra na conta

- **O modelo de voz cobra por uso.** Para a voz é obrigatório usar uma chave de API. A conexão com a conta do ChatGPT, que serve para respostas de texto sem pagar por token, não funciona para áudio.
- **Duração máxima da ligação.** Defina um teto em segundos. Isso evita chamada esquecida aberta, gastando sem ninguém do outro lado.
- **Base de conhecimento.** Cadastrar PDFs gera embeddings e exige uma chave com crédito. Base ligada direto no agente é consultada em toda resposta, o que deixa mais preciso e mais caro. O nó separado de base de conhecimento consulta só quando precisa, por exemplo quando o cliente pergunta preço.

- **Transcrição e análise.** Para transcrever as ligações e gerar a análise da IA, é preciso colocar a chave da OpenAI no Synapse.
- **A API de ligações.** O AstraCalls é open source e não cobra por ligação. A Meta também não, porque não é a API oficial. O preço disso é o risco de bloqueio do número.

Quantas ligações ao mesmo tempo

Na live eu digo que o AstraCalls não tem limite de chamadas simultâneas, e que 30 pessoas ligando ao mesmo tempo seriam 30 atendimentos da IA. A configuração da API, porém, tem um limite de ligações por sessão que você define na stack, e a escala alta ainda foi pouco testada. Os detalhes estão no artigo sobre ligações do WhatsApp no Chatwoot. Para uma operação com muito volume, teste a sua carga real antes de prometer para o cliente.

Casos de uso que combinam

O módulo de atendimentos do AstraChat foi feito para quem vive de agenda: clínicas médicas, odontológicas e de estética, barbearias e qualquer negócio que marca serviço. Os profissionais acompanham pelo aplicativo (PWA) no celular. O mesmo raciocínio vale para quem atende por texto, como mostro no artigo sobre agendamento pelo WhatsApp no Chatwoot.

Outros usos que a arquitetura permite:

- pré-atendimento por voz antes de passar para um humano;
- consulta de status de pedido ou entrega pelo nó HTTP, buscando no seu sistema;
- atendimento fora do horário, quando não tem ninguém para pegar o telefone.

Limites e cuidados

- **Só AstraCalls, por enquanto.** Na época da live, sem API oficial. Confirme o estado atual antes de planejar.
- **API não oficial.** Existe risco de bloqueio do número. Na live eu lembro que até chatbot mal construído pode levar ao banimento, não só o disparo em massa.
- **O prompt precisa dizer quando usar cada ferramenta.** A IA costuma entender sozinha, mas escrever no prompt "se o cliente quiser falar com alguém, execute o nó atribuir" deixa o comportamento mais previsível.
- **Dados sensíveis fora da conversa.** Não peça CPF, RG ou endereço pela ligação ou pelo chat. Mande um link de formulário.

- **Agenda bem configurada.** Sem serviços e disponibilidade corretos, a IA conversa bem e não agenda nada.
- **Diga que é uma assistente virtual.** Na live ela já atende assim. Quem sabe que fala com uma IA denuncia menos e tem menos frustração.

Checklist antes de ligar o agente de voz

- ☐ Ligações ativadas na caixa conectada ao AstraCalls
- ☐ Chave de API do modelo de voz com crédito
- ☐ Serviços, profissionais e horários cadastrados
- ☐ Prompt com o que a IA pode e não pode prometer
- ☐ Nó de atribuição ligado para transferir a um humano
- ☐ Duração máxima definida
- ☐ Transcrição salva no ticket e conferida em algumas chamadas
- ☐ Teste com várias ligações seguidas antes do horário de pico

Perguntas frequentes

O que é um agente de voz com IA?

É uma IA que atende a ligação e conversa com o cliente por voz em tempo real. No Flow de voz do AstraChat, ela consulta a agenda, marca horário e transfere a chamada para alguém da equipe quando o cliente pede.

Uma IA pode atender ligação do WhatsApp?

Sim. No AstraChat, o cliente liga para o número conectado pelo AstraCalls e a IA assume a chamada. Na live de setembro de 2026, o agente de voz funcionava só com o AstraCalls, e a conexão com a API oficial estava em desenvolvimento.

Quais modelos de IA servem para atendimento por voz?

Só modelos que recebem e devolvem áudio em tempo real, hoje da OpenAI e do Google Gemini. Na experiência mostrada na live, a voz da OpenAI flui melhor e o Gemini às vezes atrasa, mas a OpenAI sai mais cara.

O agente de voz com IA cobra por ligação?

O modelo de voz cobra por uso e exige uma chave de API, porque a conexão com a conta do ChatGPT não funciona para áudio. O AstraCalls é open source e não cobra por ligação, mas, por não ser a API oficial, existe risco de bloqueio do número.

O agente de voz consegue agendar horário por telefone?

Sim. No teste da live, a IA perguntou o serviço, a preferência de profissional e data, ofereceu os horários livres, pediu o nome e confirmou, e o agendamento apareceu na hora no módulo de atendimentos. Para isso, serviços, profissionais e disponibilidade precisam estar configurados.

Quer o agente de voz no seu atendimento?

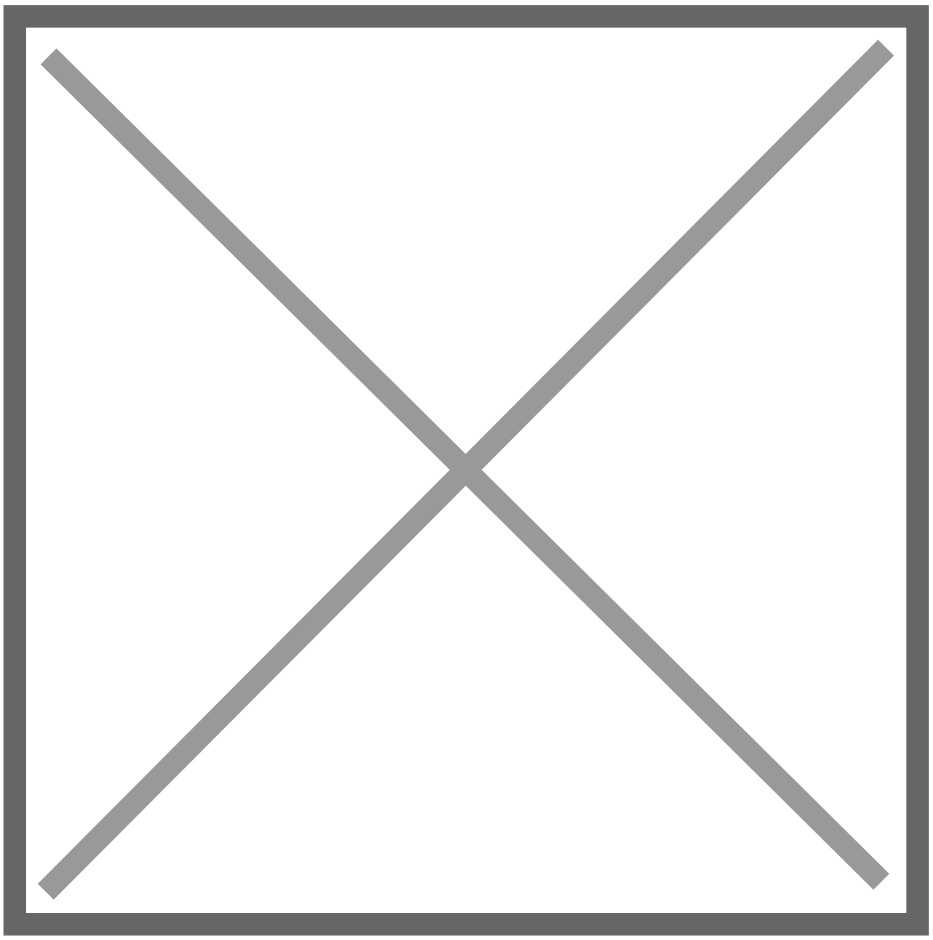
O Flow de voz faz parte do **AstraChat**, a nossa versão do Chatwoot, que traz também chatbot, sequências, workflows, Kanban, disparador e chat interno. Tudo no seu servidor, com a sua marca e com quantos atendentes você precisar.

Ele vem na Comunidade Astra Online, com código-fonte, atualizações, grupo VIP e a primeira instalação feita pelo especialista em Chatwoot Raphael Popiolek. Se você tem uma clínica ou atende muita ligação e quer ver o agente de voz no seu caso, me chame no WhatsApp com o roteiro que a sua recepção segue hoje.

Falar com um especialistaConhecer o AstraChat

Mais artigos de IA no atendimento e o Guia completo do Chatwoot, com todos os assuntos organizados.

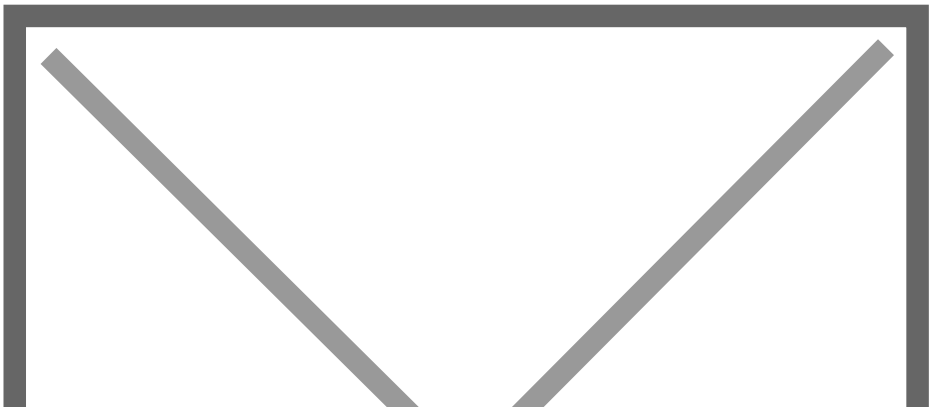
Leia também



02 de novembro de 2026

Multiatendimento WhatsApp grátis e open source: o custo real

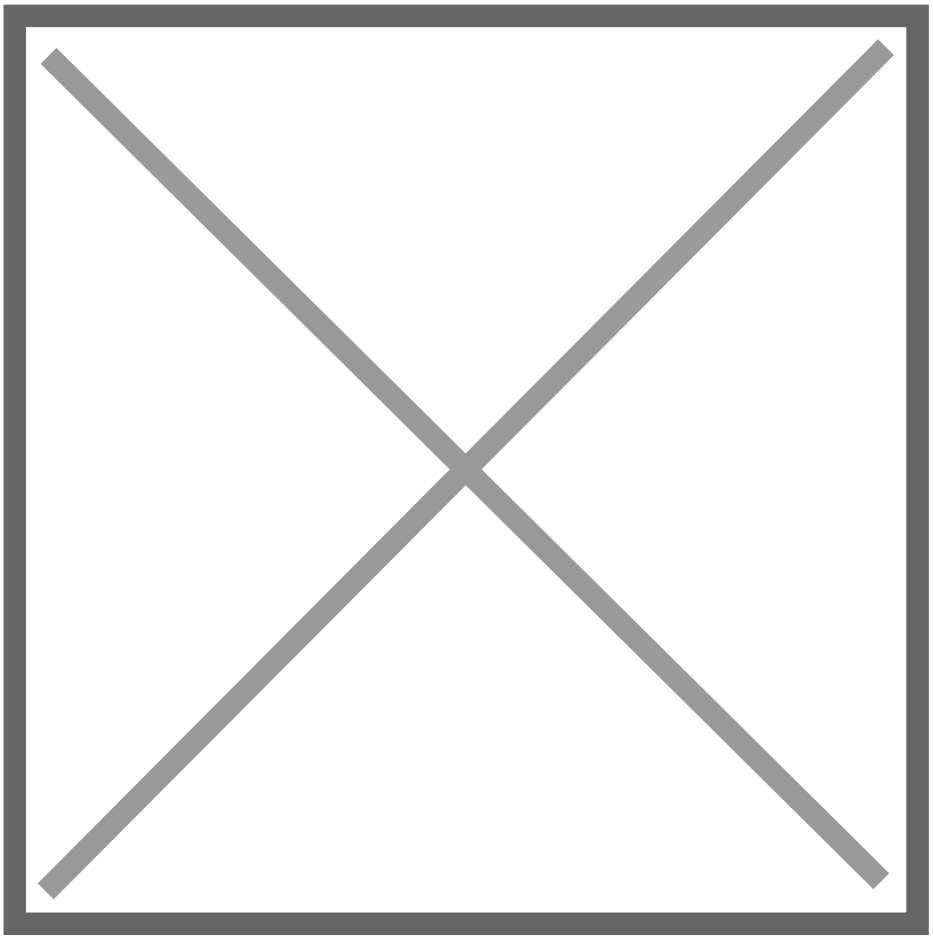
Multiatendimento WhatsApp grátis existe: Chatwoot, Whaticket, WAHA e QuePasa são open source. Veja o que dá para montar e quanto custa em servidor.



28 de outubro de 2026

Agendamento pelo WhatsApp no Chatwoot: agenda, lembretes e IA

Agendamento pelo WhatsApp dentro do Chatwoot: agenda por profissional, lembretes automáticos com data e hora e um agente de IA que marca horário sozinho.



28 de outubro de 2026

IA que atende ligação no WhatsApp: como montar a sua numa VPS

IA que atende ligação no WhatsApp e liga para a sua lista: veja como montei o painel com AstraCalls, Gemini e Claude Code numa VPS em uns 10 minutos.

© 2026 Astra Online · AstraChat · Blog · Canal no YouTube

Revision #2
Created 2026-11-09 01:08:44 UTC by Administrador
Updated 2026-11-09 01:12:51 UTC by Administrador