DOCUMENTO TÉCNICO · USO INTERNO

Assistente TBR

Assistente de atendimento por WhatsApp — especificação, custos e limitações

VERSÃO	DATA	RESPONSÁVEL TÉCNICO
1.0	01 de novembro de 2026	Jhoel Ericlis Rands

SITUAÇÃO	LOCALIZAÇÃO	SERVIÇO
Em operação · canal pendente de conexão	`/root/instinct-tbr`	`instinct.service`

> **Sobre este documento.** Todo número aqui apresentado foi **medido** no sistema em operação, não estimado. As únicas projeções são as da seção 8.2, marcadas como tal. As seções 9 e 12 registram limitações e pendências sem atenuá-las.

1. Objetivo

Este documento especifica o **Assistente TBR**, um assistente pessoal que opera por **WhatsApp** — sem aplicativo próprio, sem interface nova: o aplicativo de mensagens que o usuário já usa é a interface. O assistente guarda informações, cria lembretes com horário, registra anotações, consulta o estado do servidor e devolve resultados em linguagem natural.

O sistema foi construído para operar **exclusivamente com os recursos já existentes na infraestrutura** — sem GPU, sem contratação de serviço pago e sem tráfego de dados para fora do servidor. O objetivo declarado é servir como **porta de entrada comercial** para os serviços de automação da TBR Brasil, oferecendo a experiência de um assistente conversacional com execução real de ações sob controle do titular.

2. Escopo

2.1 O que o sistema é

- Assistente conversacional de **escopo definido** (nível 1), com ações determinísticas;
- Multiusuário, com isolamento de dados por titular;
- Com convites, consentimento granular, limites de uso e exclusão de dados;
- Com registro de auditoria de toda ação executada;
- Executado inteiramente no servidor do responsável técnico.

2.2 O que o sistema não é

- **Não é** um agente de uso geral: não navega em sites, não opera telas e não executa tarefas abertas;
- **Não realiza transações** — não compra, não paga, não reserva;
- **Não substitui** sistemas de gestão, ERP ou CRM;

- **Não utiliza** modelo de linguagem de fronteira na configuração atual (ver seção 8.2);
- **Não possui, na versão 1.0**, cobrança, autoatendimento do cliente ou autenticação no painel administrativo.

3. Descrição funcional

Todas as funções abaixo foram verificadas em execução (ver seção 10).

Função	Comportamento	Controle
Memória	Guarda fatos informados pelo usuário, com deduplicação automática; permite consulta e esquecimento pontual.	Requer permissão `memoria`
Lembretes	Interpreta expressões de tempo em português ("amanhã às 10h", "em 20 minutos", "sexta às 8h") e avisa o usuário no momento devido, sem intervenção.	Requer permissão `lembretes`
Anotações	Cria e pesquisa anotações junto com a memória.	Requer permissão `anotacoes`
Estado do servidor	Informa uso de disco, memória disponível, carga do processador e situação dos serviços monitorados.	Requer permissão `servidor`
Áudio	Mensagens de voz são transcritas localmente, sem serviço externo.	Herdado do canal
Convites	Cada usuário dispõe de 10 códigos de convite; link público, de uso único.	—
Consentimento	O titular liga e desliga cada recurso; o assistente recusa a ação quando desligado.	—
Limites de uso	Teto diário de mensagens e de consultas à IA, por usuário.	Administrador
Exclusão de dados	Apaga integralmente os dados do titular, inclusive a mensagem de confirmação.	Exige confirmação textual
Auditoria	Registra ferramenta, entrada, saída, situação e tempo de cada ação.	—
Painel administrativo	Usuários, conversas, auditoria, conhecimento, convites e eventos.	Sem autenticação (ver 9.5)

4. Arquitetura

O sistema organiza-se em cinco camadas. A **camada de atuação** é deliberadamente restrita a operações determinísticas — é essa restrição que torna a confiabilidade alta e o custo marginal nulo.

Camada	Componentes	Tecnologia
1. Canal	Recepção de webhook, normalização de texto, áudio e imagem, envio de resposta	Evolution API (WhatsApp)
2. Cérebro	Rota determinística por padrões; rota por modelo de linguagem com saída restrita a JSON; histórico; sessão	Python · Ollama (qwen2.5:3b)
3. Atuação	13 ferramentas determinísticas com verificação de permissão	Biblioteca padrão do Python
4. Dados	Usuários, convites, permissões, uso, mensagens, memória, lembretes, anotações, auditoria, eventos	SQLite (WAL)
5. Controle	Limites diários, consentimento, exclusão, agendador de lembretes e vigilância de disco	Python · systemd

4.1 Decisão de projeto: duas rotas, não uma

Mensagens que correspondem a padrões conhecidos são resolvidas por **regras determinísticas**, com tempo de resposta de milissegundos e comportamento previsível. Somente o que não corresponde a nenhum padrão é encaminhado ao modelo de linguagem, com a resposta restringida a um formato JSON fixo. Se o modelo devolver conteúdo inválido, o sistema **solicita reformulação** em vez de executar uma ação incerta.

A consequência prática é que **a rota crítica não depende do modelo**: mesmo com o modelo de linguagem indisponível, as funções essenciais continuam operando.

5. Modelo de dados

Tabela	Conteúdo
`usuarios`	Telefone (identificador), nome, plano, situação, limites individuais
`convites`	Código, quem convidou, quem usou, quando
`permissoes`	Consentimento por recurso, por titular
`uso`	Mensagens e consultas de IA por dia, por usuário
`mensagens`	Histórico de conversa (papel, texto, canal)
`memoria`	Fatos declarados pelo titular
`lembretes`	Texto, expressão original, horário calculado, situação de aviso
`anotacoes`	Título e conteúdo
`acoes`	Auditoria: ferramenta, entrada, saída, situação, duração
`eventos`	Marcos do sistema: novo usuário, convite usado, dados apagados, avisos

Isolamento: todas as consultas são filtradas pelo telefone do titular. Não existe consulta que atravesse usuários nas rotas de atendimento.

6. Segurança e privacidade

- **Confinamento de rede:** o serviço escuta exclusivamente em `127.0.0.1:8020`. Exposição externa exige configuração deliberada no proxy reverso.
- **Sem tráfego externo:** na configuração atual, nenhuma mensagem ou dado sai do servidor. A transcrição de áudio e a inferência ocorrem localmente.
- **Credenciais:** a chave da API do canal é lida do contêiner em tempo de execução e não é gravada em arquivo, código ou registro.
- **Consentimento verificável:** a checagem de permissão ocorre no momento da execução, dentro do componente que executa a ação — não na interface.
- **Auditoria:** toda ação é registrada com entrada, saída, situação e duração.
- **Direito ao esquecimento:** a exclusão remove os dados do titular de todas as tabelas, inclusive a mensagem que confirma a exclusão.

> **Ponto de atenção jurídica.** A partir do primeiro atendimento a terceiros, o responsável passa a tratar dados pessoais de titulares alheios, submetendo-se à LGPD. As ferramentas de consentimento, limitação de finalidade e exclusão existem e funcionam; a definição de base legal, retenção e encarregado é decisão do responsável, não do software.

7. Operação

Item	Valor medido
Gerência do processo	`systemd` — `instinct.service`, habilitado no boot
Política de reinício	`Restart=always`, intervalo de 5 s
Encerramento abrupto (`kill -9`)	**Recuperação automática verificada** — serviço restabelecido e escutando
Registro de execução	`/root/instinct-tbr/estado/servico.log` (append)
Migração de banco	Automática na inicialização (colunas e índices)
Monitoramento próprio	Agendador verifica lembretes a cada 30 s e disco a cada 6 h

```
systemctl status instinct.service
systemctl restart instinct.service
journalctl -u instinct.service -f
tail -f /root/instinct-tbr/estado/servico.log
```

8. Custos

8.1 Custo corrente da configuração atual (medido)

Item	Consumo medido	Custo marginal
Inferência de linguagem	Modelo local `qwen2.5:3b`, 100% em CPU	**R$ 0,00**
Transcrição de áudio	Modelo local `faster-whisper small`	**R$ 0,00**
Armazenamento do projeto	1,5 MB (código + páginas)	**R$ 0,00**
Banco de dados	4 KB com 4 usuários e 34 mensagens	**R$ 0,00**
Memória do serviço	28 MB	Reaproveita servidor existente
Modelo residente em memória	**2,2 GB** continuamente	Custo de oportunidade

O custo marginal de operação é **nulo**: o sistema não consome serviço pago e reaproveita o servidor e o canal de mensagens já existentes. O custo real é o **de oportunidade**: 2,2 GB de memória ficam permanentemente alocados ao modelo, em um servidor com 7,8 GB no total.

8.2 Projeção com modelo de linguagem remoto (estimativa)

A configuração atual pode ser substituída por um modelo remoto alterando **uma variável de ambiente**. O custo passa a ser por uso. A base de cálculo são os tokens **medidos** em uma chamada real de roteamento:

Grandeza	Valor medido
Tokens de entrada (prompt de roteamento)	548
Tokens de saída (resposta JSON)	25
Chamadas ao modelo por mensagem fora do padrão	1
Mensagens que exigem o modelo (observado em teste)	1 em cada 13

Aplicando os preços publicados de duas faixas de modelo, apenas sobre as mensagens que exigem o modelo:

Faixa de modelo	Por 1.000 mensagens roteadas	Volume de 30.000 mensagens/mês
Econômico (classe 0,3 / 1,1 US$ por milhão)	≈ US$ 0,18	≈ US$ 5,26
Fronteira (classe 3,0 / 15,0 US$ por milhão)	≈ US$ 2,02	≈ US$ 60,57

> **Ressalvas obrigatórias desta projeção.** (i) Os preços unitários são de referência e devem ser confirmados na data da contratação. (ii) O cálculo pressupõe uma única chamada por mensagem, sem tentativas repetidas. (iii) Não estão incluídos áudio, imagens, histórico longo nem memória semântica, que elevariam a contagem de tokens. (iv) A configuração atual **não incorre em nenhum destes custos**.

8.3 Custos não monetários

- **Latência:** a rota por modelo responde em 6,8 a 9,2 segundos, contra 0,01 a 0,03 segundo da rota determinística. É o preço de não usar GPU.
- **Memória:** 2,2 GB permanentemente alocados ao modelo residente.
- **Disco:** o servidor opera com 93,1% de ocupação (6,6 GB livres) — **restrição pré-existente e crítica**, não causada por este sistema, que ocupa 1,5 MB.

9. Limitações

Limitação	Consequência prática	Natureza
Atuação restrita ao nível 1	Não navega em sites, não opera telas, não executa tarefas abertas. Apenas ações previstas e determinísticas.	Arquitetural
Modelo de 3 bilhões de parâmetros	Falha em ambiguidades de linguagem. Casos já corrigidos: confusão entre "lembrar um fato" e "criar lembrete"; tratamento de "anotação" como verbo. Mitigação: ampliação das regras determinísticas.	Técnica
Capacidade de processamento	2 núcleos. Suporta poucos usuários simultâneos; chamadas ao modelo não são paralelizadas.	Infraestrutura
Painel administrativo sem autenticação	**Não expor na internet.** Qualquer acesso à porta 8020 concede controle total.	Segurança — pendência
Ausência de cobrança	Não há assinatura, faturamento ou controle financeiro do cliente.	Comercial
Dependência do canal	Sem o Evolution API conectado, o assistente não recebe nem envia mensagens. O restante do sistema permanece funcional.	Externa
Ausência de autoatendimento do cliente	O titular vê seus dados apenas pelo assistente; não há portal próprio.	Produto
Ponto único de falha	Um único servidor, sem redundância nem backup automatizado do banco.	Infraestrutura
LGPD	Base legal, prazo de retenção e encarregado não definidos. Responsabilidade do controlador.	Jurídica

10. Evidência de aceitação

Os resultados abaixo foram produzidos por execução automatizada contra o serviço em operação, pelo mesmo caminho utilizado pelo canal de mensagens.

> **Bateria de produto — 30 de 30 verificações aprovadas.**
> Cobre: emissão e resgate de convite, uso único do convite, recusa de convite já utilizado, boas-vindas ao novo usuário, contabilização de uso, cálculo de horário do lembrete, execução do

> agendador, preservação do lembrete quando o envio falha, desligamento e religamento de permissão, bloqueio efetivo da ação sem permissão, configuração e aplicação de limite de uso, exclusão de dados com confirmação, remoção integral dos dados e registro do evento de exclusão.

> **Bateria de roteamento — 13 de 13 casos corretos (100%).**
> 12 casos resolvidos pela rota determinística e 1 pela rota de modelo, com acerto integral.

> **Resistência a falha — verificada.**
> O processo foi encerrado abruptamente com `kill -9`. O serviço foi restabelecido automaticamente pelo `systemd` e voltou a escutar na porta configurada.

> **Degradação controlada — verificada.**
> Com o canal de mensagens desconectado, o assistente processa a entrada, executa a ação, registra a auditoria e falha **somente no envio**, sem interrupção do serviço e sem perda do lembrete.

10.1 Defeitos identificados e corrigidos durante o desenvolvimento

1. **Boas-vindas não era enviada** a quem ingressava por convite, pois o registro do usuário já existia antes da primeira mensagem. Correção: o critério passou a ser "nunca trocou mensagem".
2. **Agendador não deixava rastro** quando o envio falhava, impossibilitando comprovar sua execução. Correção: a tentativa é registrada na auditoria, e o lembrete **não** é marcado como avisado, garantindo nova tentativa.
3. **Exclusão de dados deixava resíduo** — a mensagem de confirmação era gravada após a remoção. Correção: nenhum registro é mantido após a exclusão, inclusive a confirmação.
4. **Interpretação incorreta de "depois de amanhã"**, resolvida como "amanhã". Correção: ordem de avaliação das regras temporais.
5. **Ausência de migração de banco**, impedindo a atualização sobre base existente. Correção: migração automática de colunas e índices na inicialização.

11. Dimensão do sistema

Item	Quantidade
Código de aplicação (Python)	1.665 linhas
Verificação automatizada (testes e demonstrações)	503 linhas
Instrumentos de reconhecimento	48 linhas
Total em Python	**2.216 linhas**
Interfaces web (HTML/CSS/JS, sem dependências)	404 linhas
Bibliotecas externas instaladas	0
Dependências obrigatórias	Python 3.11, SQLite, systemd

O sistema não utiliza nenhum pacote de terceiros além dos já presentes na infraestrutura. Toda a camada de rede, roteamento, persistência, interface e servidor HTTP apoia-se na biblioteca padrão do Python, reduzindo a superfície de manutenção e de vulnerabilidade.

12. Pendências e próximos passos

12.1 Dependem do responsável

Pendência	Efeito	Esforço
Reconectar a instância do WhatsApp (leitura de QR) e apontar o webhook para `/webhook/evolution`	Sem isso o assistente está operacional mas não recebe mensagens	Minutos
Definir o modelo de linguagem: local (custo zero, maior latência e imprecisão) ou remoto (melhor compreensão, custo por uso)	Determina custo e qualidade da experiência	Minutos
Proteger o painel com autenticação antes de exposição externa	Mitiga risco de acesso indevido	Horas
Liberar espaço em disco (93,1% ocupado)	Risco de indisponibilidade dos serviços do servidor	Horas
Definir base legal e retenção (LGPD)	Requisito para operação comercial	Externo

12.2 Evolução técnica sugerida, em ordem de retorno

Ordem	Item	Ganho	Esforço estimado
1	Substituição do modelo local por modelo remoto de melhor qualidade	Compreensão de linguagem natural; queda acentuada de erro de interpretação	1 dia
2	Ferramenta de negócio real (ex.: leitura de documentos guiada por checklist)	Transforma o assistente em produto específico, com valor demonstrável	1 a 2 semanas
3	Memória semântica e integração com correio eletrônico e agenda	Aproximação da experiência de "assistente que conhece o usuário"	2 a 4 semanas
4	Portal de autoatendimento do cliente e cobrança	Permite comercialização direta	4 a 8 semanas
5	Atuação por navegador (nível 2)	Execução em sistemas sem interface programática	3 a 6 semanas + custo por tarefa

> O item 5 e qualquer evolução para atuação geral (nível 3) **não são viáveis na infraestrutura atual**: exigem modelo de fronteira, unidade de processamento gráfico e equipe dedicada. A restrição não é de software, mas de capacidade computacional e de capital.

13. Conclusão

O Assistente TBR encontra-se **operacional e verificado** na configuração descrita. Entrega, com confiabilidade comprovada, as funções de memória, lembretes com aviso automático, anotações, consulta ao estado do servidor, convites, consentimento, limites de uso, exclusão de dados e auditoria — **a custo marginal nulo** e sem tráfego de dados para fora do servidor.

A pendência que separa o sistema de sua operação efetiva é **externa ao software**: a reconexão do canal de mensagens. As demais limitações estão identificadas na seção 9 e são, em sua maioria, de natureza comercial e jurídica, não técnica.