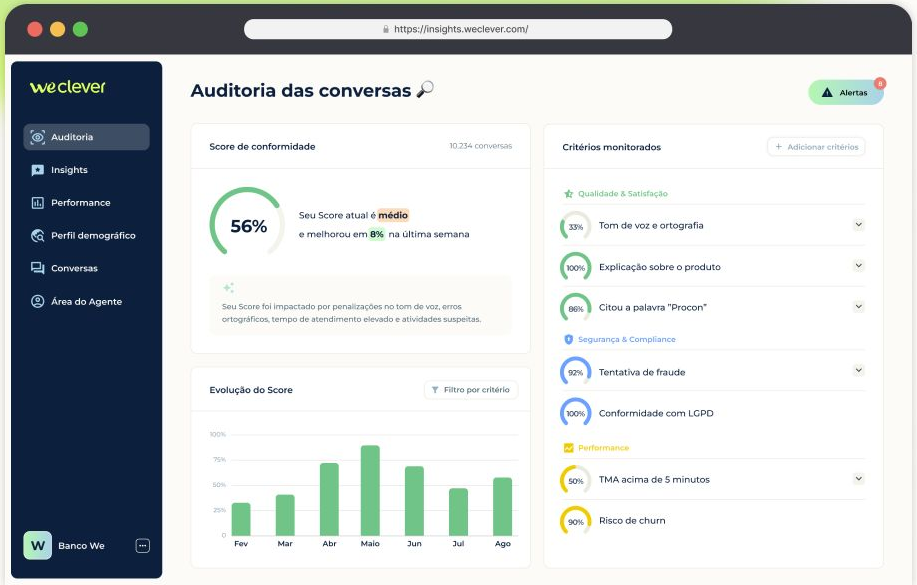
Primeiro Meetup da Welcever

Bem vindos DEVs!

Primeiro M🙂o WeClever

O que é a WeClever !?

Plataforma de Inteligência Conversacional



A WeClever simplifica o complexo

Produto: Conversas & Jornadas

Mensageria
- Integração oficial com WhatsApp Business API
- Compra, portabilidade e gestão de qualidade dos números
- Plataforma de conversa para transbordo

Jornada
- Desenho e otimização das etapas de atendimentos
- Reengajamento e reaquecimento de leads
- Transbordo para humanos

Inteligência Artificial
- Integração com LLMs e APIs
- Transcrição de áudio e imagem
- Prompt engineering e Fine-tuning

Produto: Auditoria & Monitoria

Auditoria
- Score de assertividade da conversas
- Monitoramento de qualidade das conversas
- Atuação em tempo real em atendimentos detratores

Insights
- Sentimento, satisfação e objeções dos clientes
- Resumo da conversa e insights acionáveis via IA
- Indicadores de performance e da jornada de atendimento

Segurança
- Compliance com regulamentações e políticas de segurança
- Detecção e mitigação de riscos de forma proativa
- Dados criptografados e em ambiente seguro

Serviços Personalizados

Performance
- Key Account Manager dedicado
- Relatórios personalizados
- Gestão de resultados e estratégias de growth

Experiência
- Conversas fluidas e hiperpersonalizadas
- Customização dos Agentes de IA
- Consultores de vendas sob demanda

Empresas líderes em seus segmentos confiam no poder da nossa Inteligência Conversacional

Bancos e Fintechs

Varejo e E-commerce

Educação

Serviços

Vamos lá 🙂

Primeiro Meetup WeClever

Objetivo do Meetup

- **Explorar Tecnologias:** Apresentar soluções com exemplos práticos.
- **Networking e Troca:** Oportunidade para desenvolver conexões e compartilhar experiências entre desenvolvedores locais.
- **Fomentar a Comunidade:** Estimular a colaboração e fortalecer a rede de desenvolvedores na região.

Como Usamos Serverless com Lambda na WeClever

André de Faria Carvalho
Desenvolvedor Backend
Cientista da computação

weclever

PUC Minas

Serverless

Sem servidor

O conceito Serverless é uma forma de executar código na nuvem sem precisar gerenciar servidores, **pagando apenas pelo que você usa**.

- AWS Lambda: Serviço central para execução de código em resposta a eventos.
- API Gateway: Criação e gerenciamento de APIs RESTful.
- S3: Armazenamento de objetos com alta durabilidade.
- DynamoDB: Banco de dados NoSQL gerenciado.
- Step Functions: Orquestração de funções Lambda.
- SQS e SNS: Fila de mensagens e notificações.

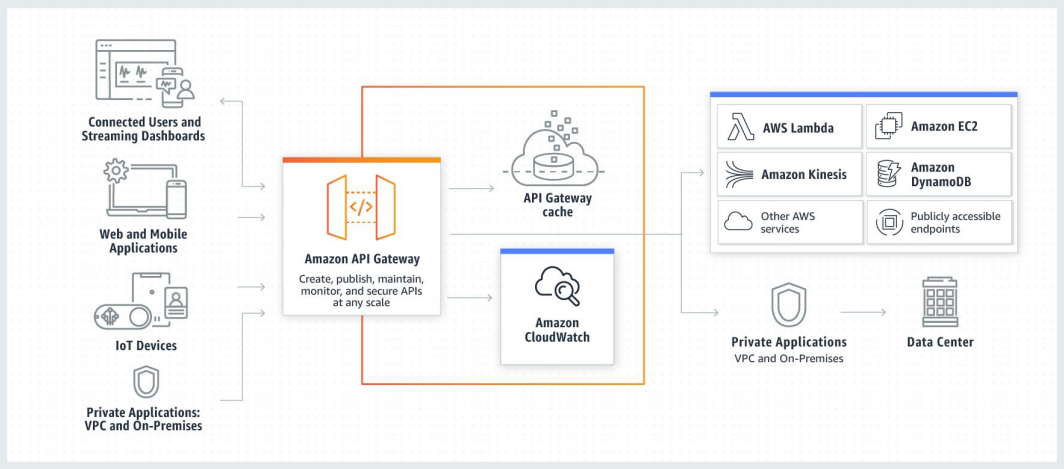
Lambda λ

AWS Lambda é um serviço de computação serverless da AWS que permite **executar código** em resposta a eventos sem gerenciar servidores.

- Escalabilidade
- Custo-eficiência
- Gerenciamento
- Escalabilidade Automática
- Integração com outros serviços da AWS

API Gateway ➡️

API Gateway é a **porta de entrada** para suas APIs (aplicações), gerenciando o tráfego, segurança e roteamento das requisições



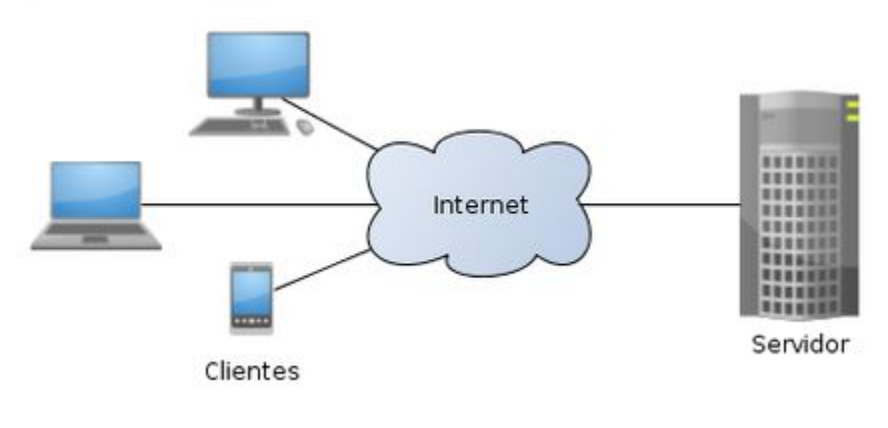
API Gateway + Lambda

API Gateway e AWS Lambda são dois serviços que, juntos, formam uma arquitetura poderosa

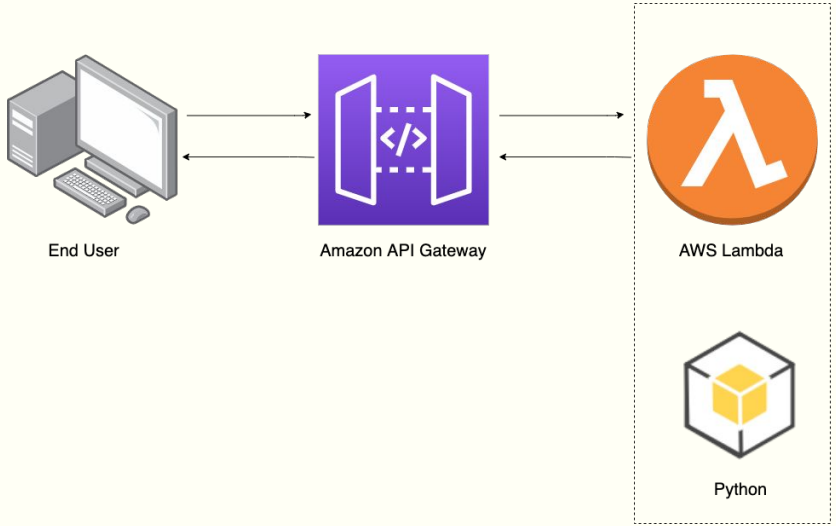
- Processamento de dados em tempo real.
- APIs sem servidor.
- Funções em resposta a eventos.
- Integração, Webhooks, APIs externas

Serverless Vs Cliente Servidor

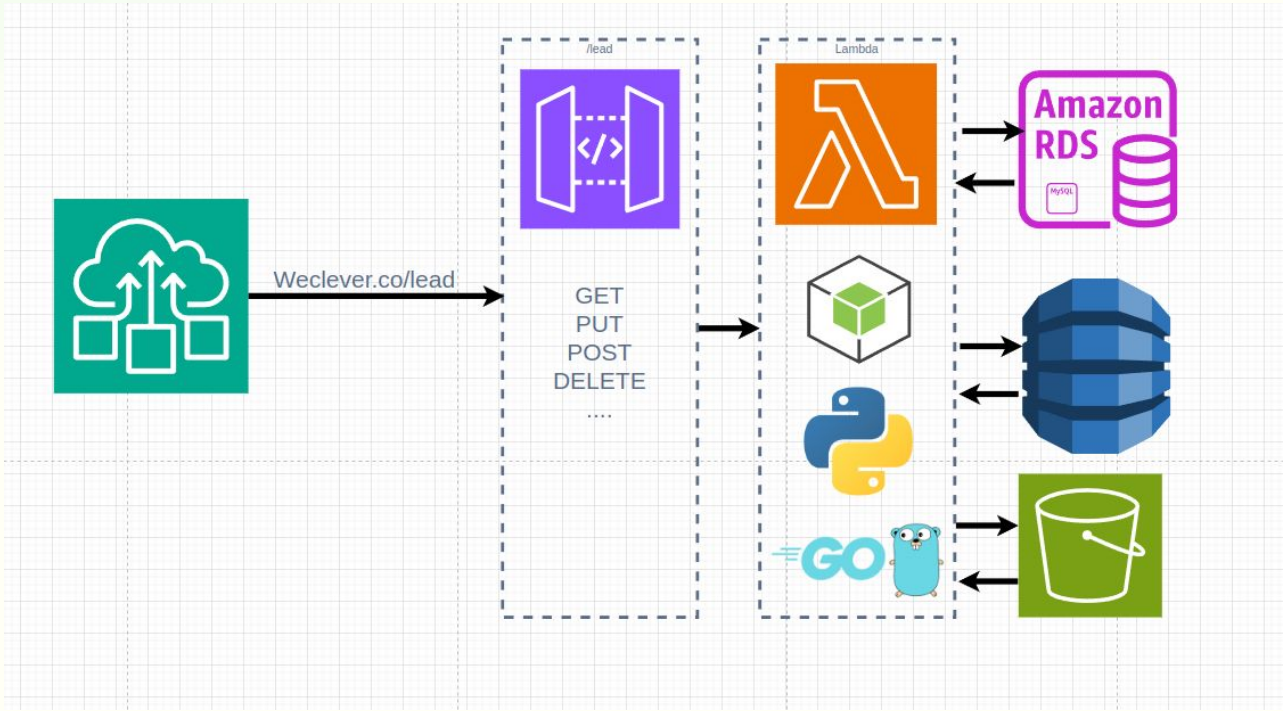
Cliente Servidor



Serverless



Arquitetura básica



Bora pro Código

Serverless Framework

Inicializando

```
serverless create --template aws-nodejs --path hello-express
cd hello-express
npm init -y
npm install express serverless-http

hello-express/
├── app.js
├── handler.js
└── serverless.yml
```

app.js

```
const express = require('express');
const serverless = require('serverless-http');

const app = express();

app.get('/hello', (req, res) => {
  res.json({ message: 'Hello, World!' });
});

module.exports.handler = serverless(app);
```

Serverless Framework

Serverless.yml

```yaml
service: hello-world

provider:
  name: aws
  runtime: nodejs18.x

functions:
  app:
    handler: serverless-http.handler
    events:
      - http:
          path: hello
          method: get
```

Executando

```
serverless deploy

curl https://xxxxxxx.execute-api.region.amazonaws.com/dev/hello
  Response: {
      "message": "Hello, World!"
  }

🎉🎉🎉 Sucesso 🎉🎉🎉
```

Pausa
10 minutos

we clever

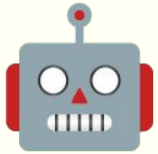
A Inteligência Artificial no Agente Conversacional da WeClever

João Pedro Rosa

Desenvolvedor Backend
Engenharia da computação

weclever

INSTITUTO FEDERAL
Sul de Minas Gerais

Agente de IA 

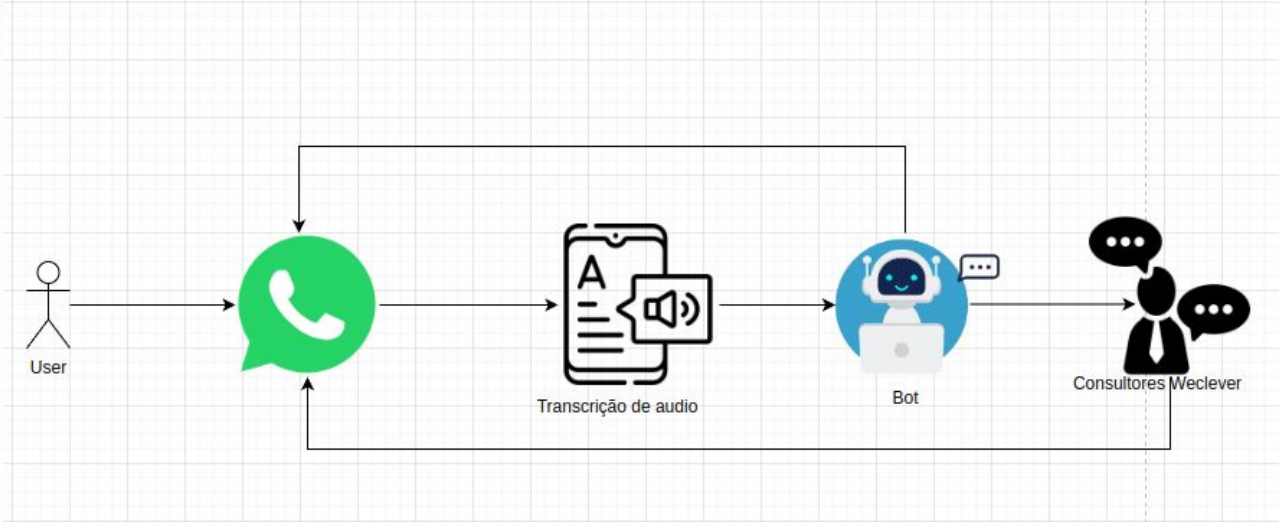
Agentes de IA são sistemas autônomos que tomam decisões e executam tarefas com base em **dados e regras (prompts)**, podendo aprender e melhorar com o tempo.

Tecnologias

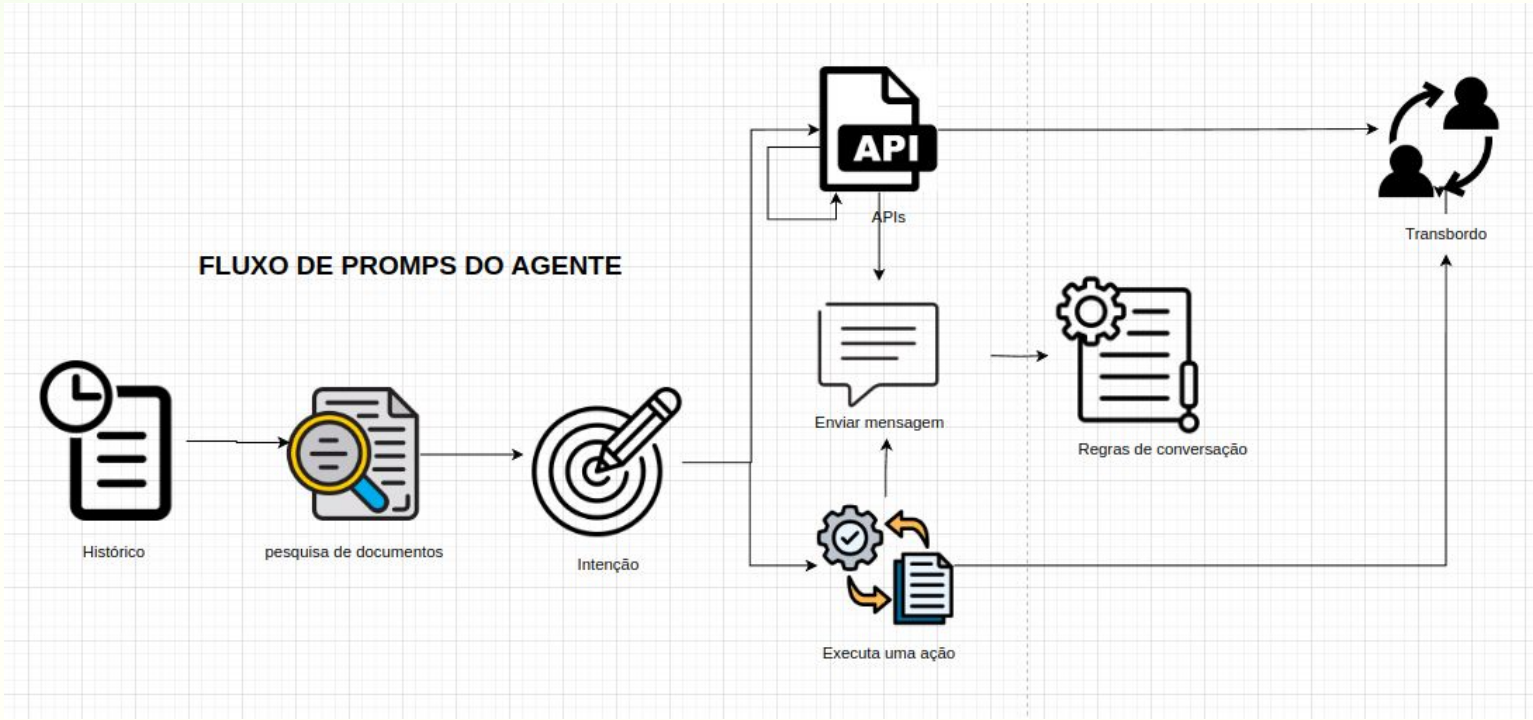
Usamos as seguintes tecnologias nos nossos agentes

- Aws Lambda
- S3
- Python e NodeJs
- Banco de dados relacional Mysql
- Banco Não relacional AWS Dynamo DB

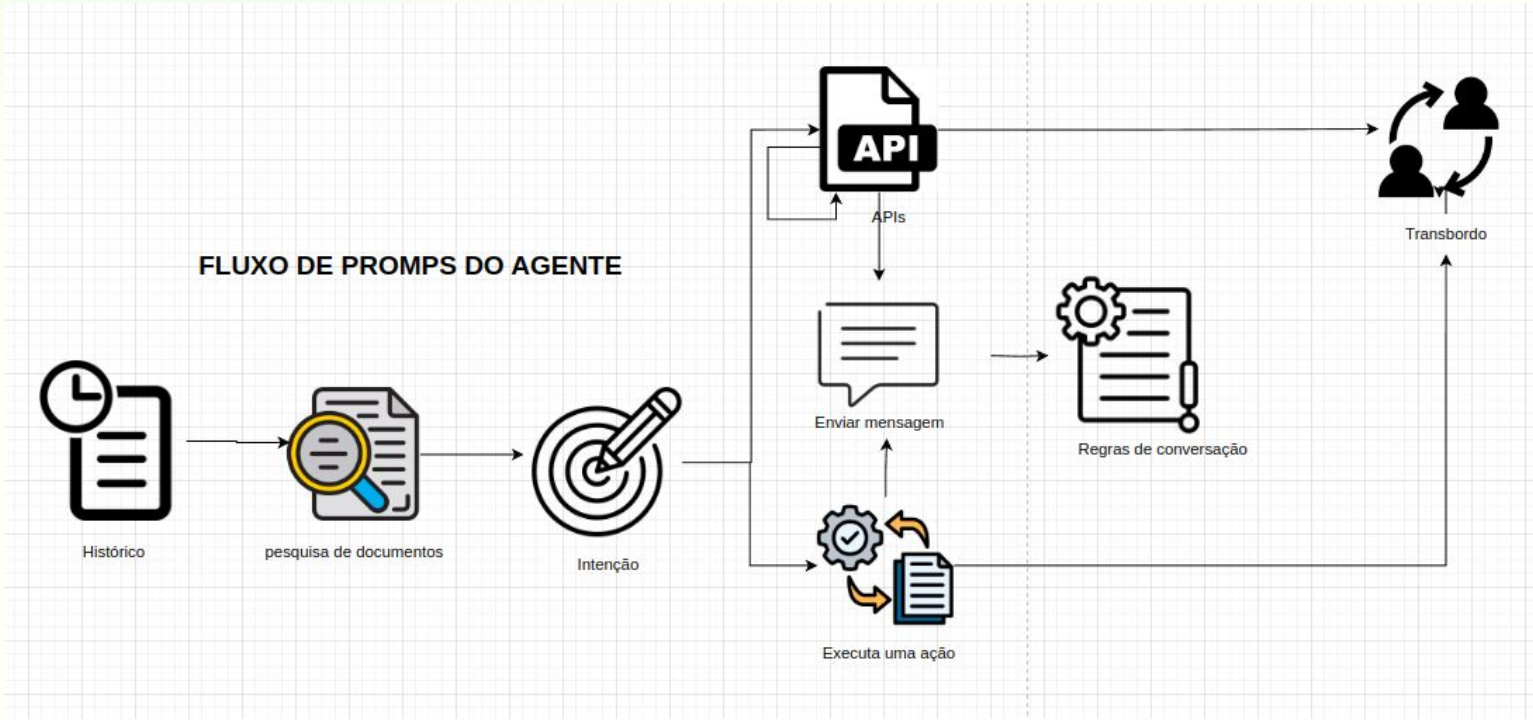
Visão Geral



Estrutura de Prompts

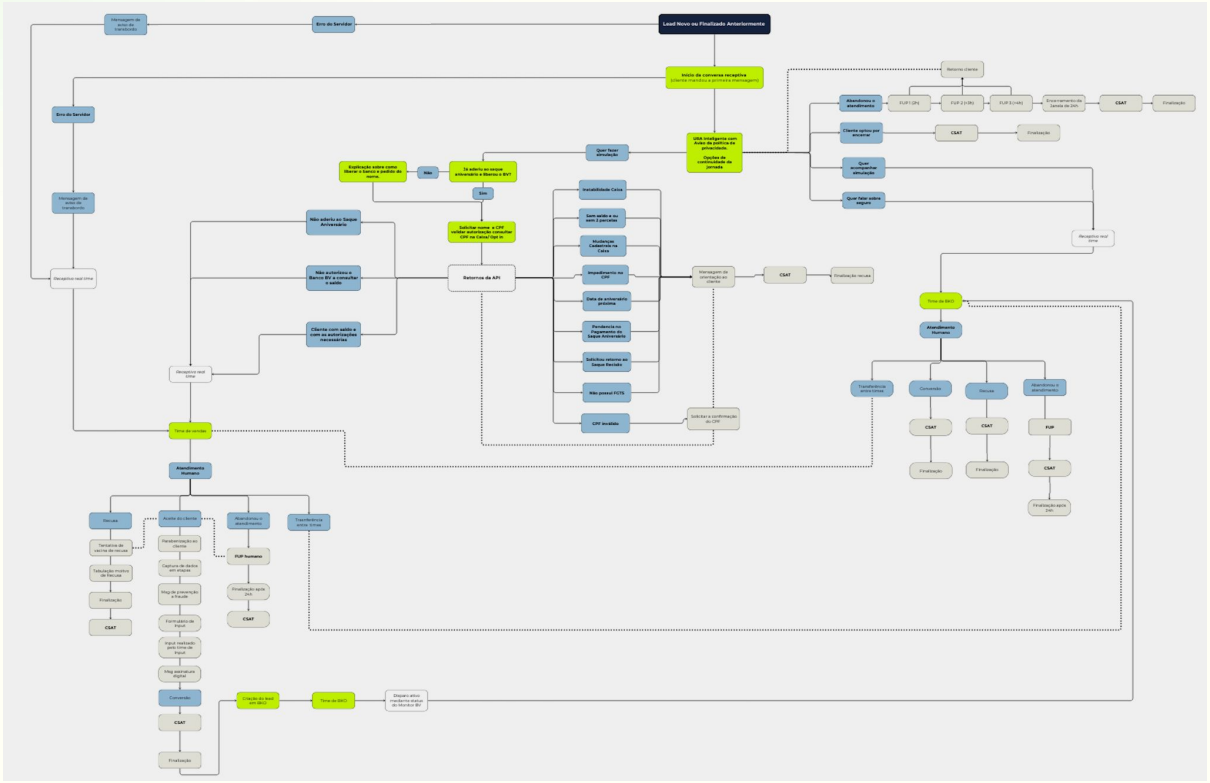


Prompts


FLUXO DE PROMPS DO AGENTE
API
APIs
Transbordo
Enviar mensagem
Regras de conversação
Histórico
pesquisa de documentos
Intenção
Executa uma ação

Teste nosso agente

Jornada complexa



DESAFIO 1 - CRIAÇÃO DE UMA JORNADA DE CONVERSA COMPLEXA

JORNADA EM LINGUAGEM NATURAL

CAPTURA DE DADOS AO LONGO DA INTERAÇÃO

ORIENTAÇÃO AO CLIENTE, FLUXO DE OPT IN E POLÍTICA DE SEGURANÇA

DEFINIÇÃO DE ELEGIBILIDADE PARA O SAQUE FGTS

BUSCA POR AUTO SERVIÇO, AUMENTANDO A PRODUTIVIDADE DA OPERAÇÃO

2.1 - Qual LLM utilizar?

GPT 3.5 TURBO

Mais barato, porém ”pouco comunicativo”

vs.

GPT - 4

Extremamente ”inteligente”, porém muito caro

Solução Inicial:

GPT 3.5 TURBO → para tomada de decisão

GPT - 4 → para conversação

Com o passar do tempo...

OpenAI GPT-4o

Bom custo-benefício!

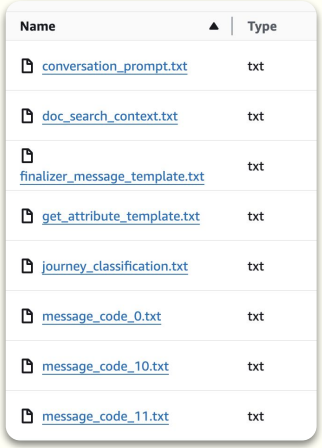
DESAFIO 2 - MANTER CONVERSA NATURAL

2.2 - PROMPT ENGINEERING

- Cada detalhe conta
- Regule a temperatura
- Evite um prompt muito extenso
- Foque no que a IA deve fazer, e não no que não pode fazer

2.3 - PROMPT ARCHITECTURE

RAG (ou geração aumentada via recuperação) →

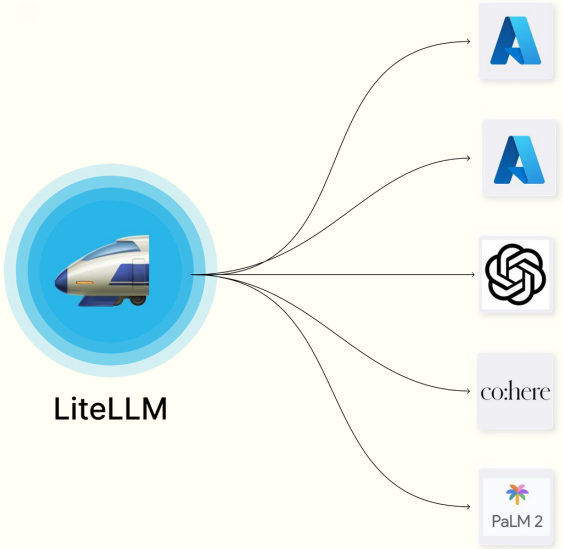


DESAFIO 2 - MANTER CONVERSA NATURAL

2.4 - STACK EXTERNA

LiteLLM: mantendo sua aplicação "LLM Agnostic"

Tokenizer: acompanhe seus custos!

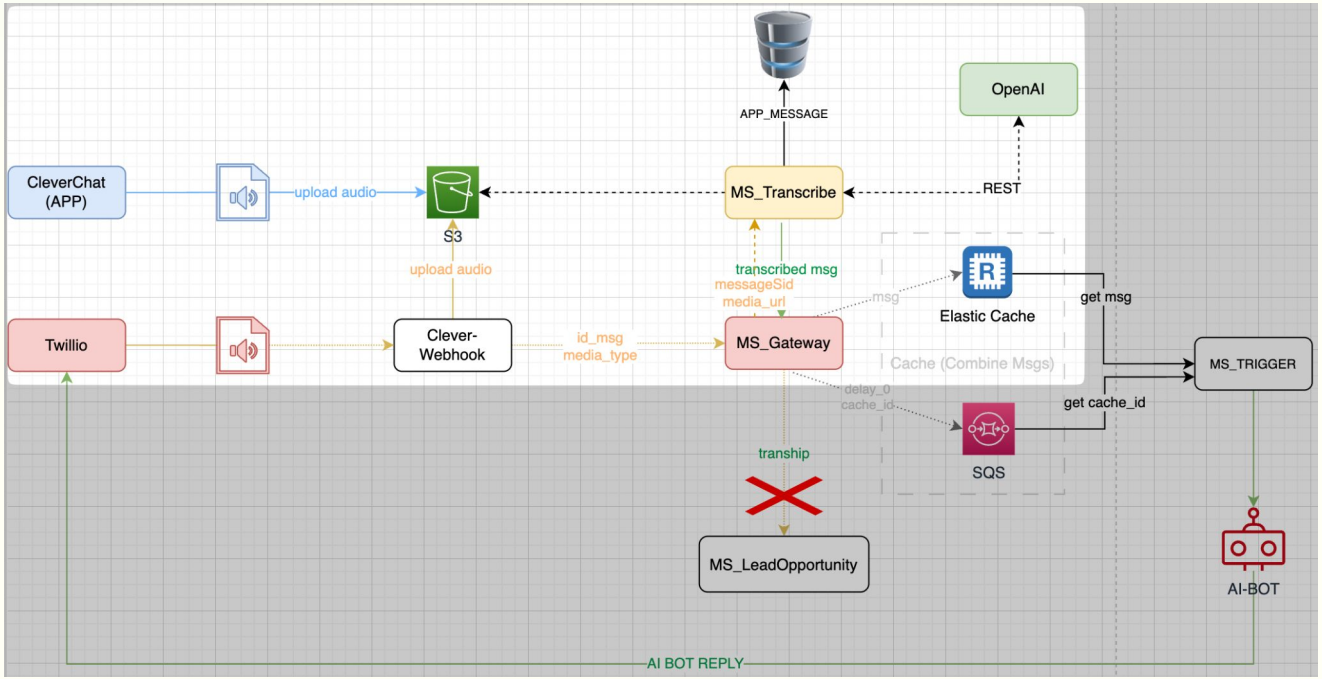


DESAFIO 2 - MANTER CONVERSA NATURAL

2.5 - E SE O CLIENTE MANDAR ÁUDIO? E SE MANDAR IMAGEM?

WHISPER
Transcrição de áudio para texto

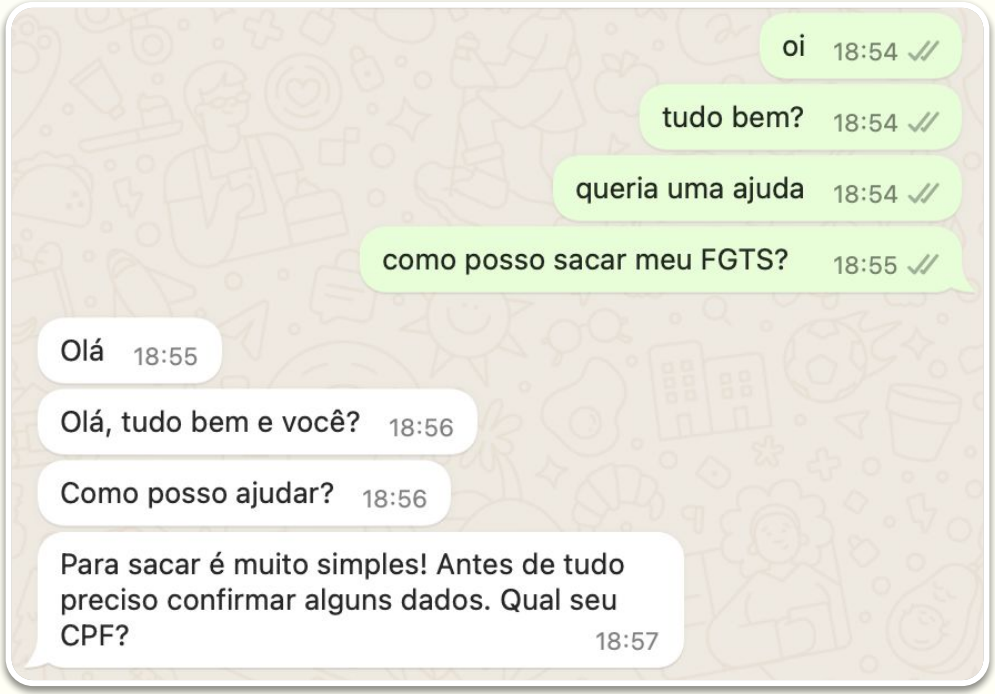
VISION
Interpretação de imagens como texto



DESAFIO 2 - MANTER CONVERSA NATURAL

2.6 - E SE O CLIENTE ESCREVER ASSIM?

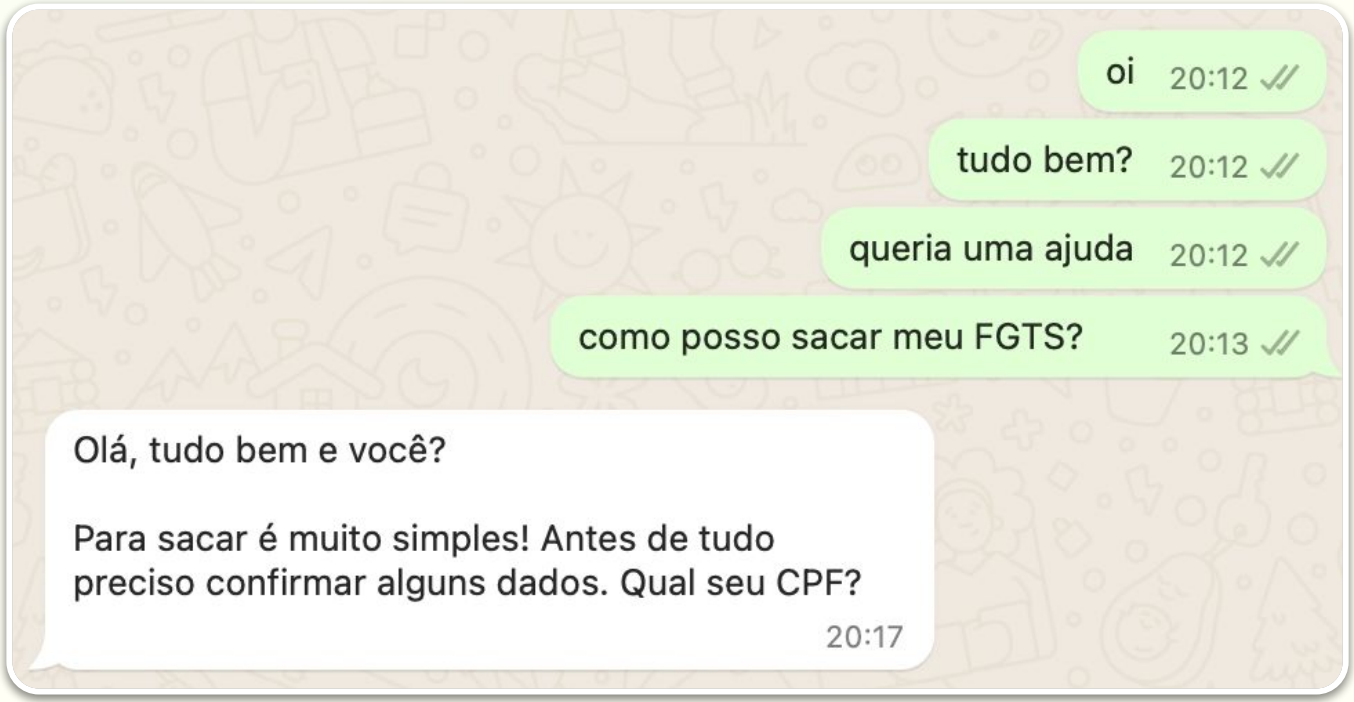
Solução ”old school”:
responder as mensagens uma a uma



DESAFIO 2 - MANTER CONVERSA NATURAL

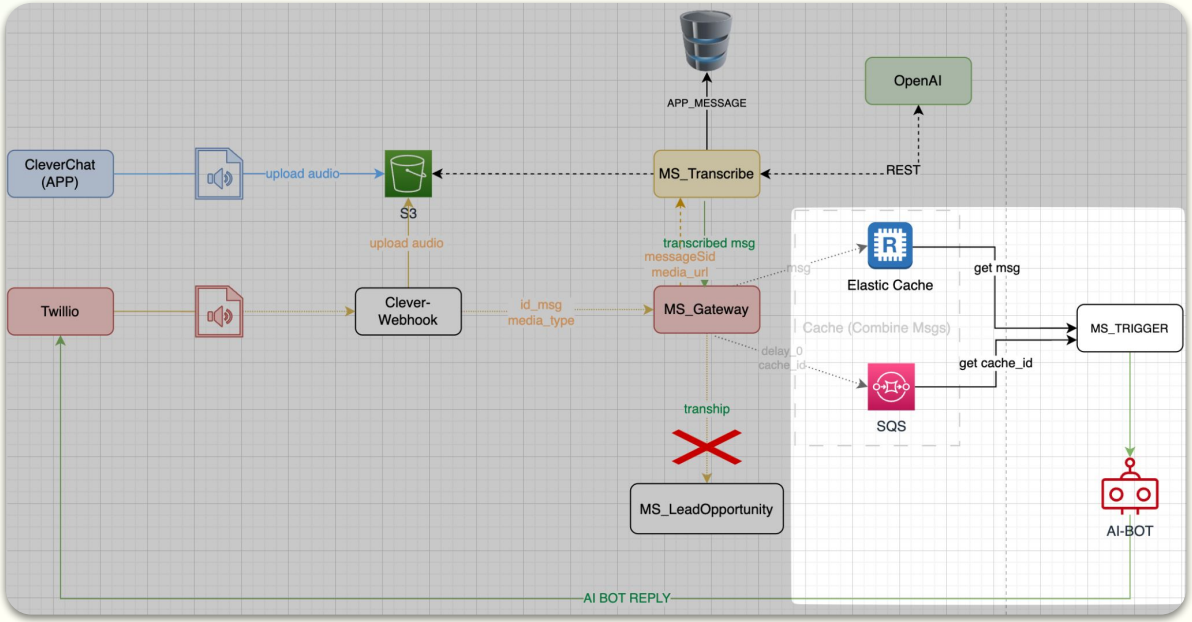
2.6 - E SE O CLIENTE ESCREVER ASSIM?

Solução ideal:



DESAFIO 2 - MANTER CONVERSA NATURAL

2.6 - E SE O CLIENTE ESCREVER ASSIM?



DESAFIO 4 - GUARD RAILS / ALUCINAÇÃO

Como fazemos nosso fine tuning:

Base com mínimo de 100 conversas auditadas e validadas como modelos a serem seguidos pelos agentes;

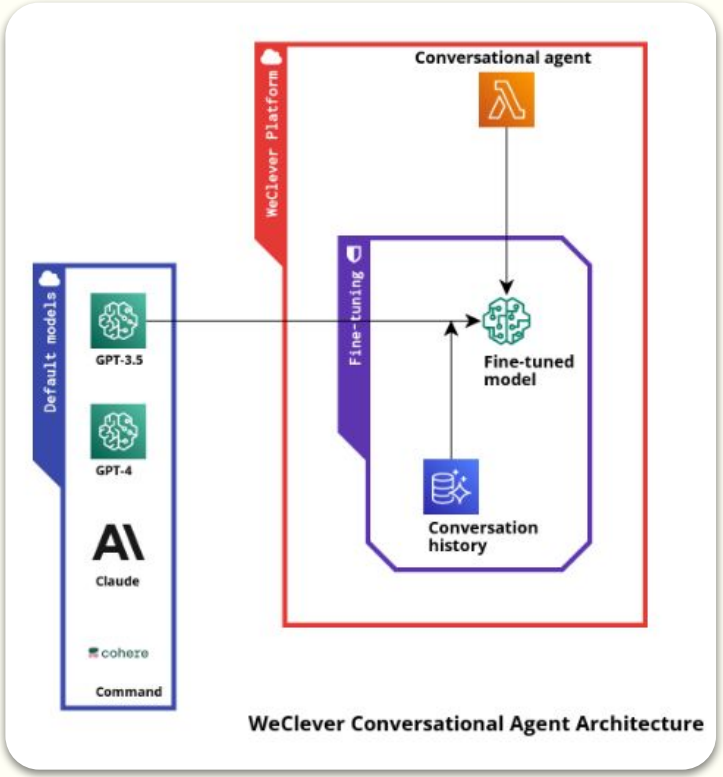
Utilizar modelos de conversas que percorreram a jornada conversacional ideal, mas também que foram abortadas pelo usuário e/ou não chegaram até o fim;

Considerar conversas que terminaram tanto em conversão quanto em recusa;

Estruturar as conversas para que sejam aproveitadas pela IA em sua totalidade;

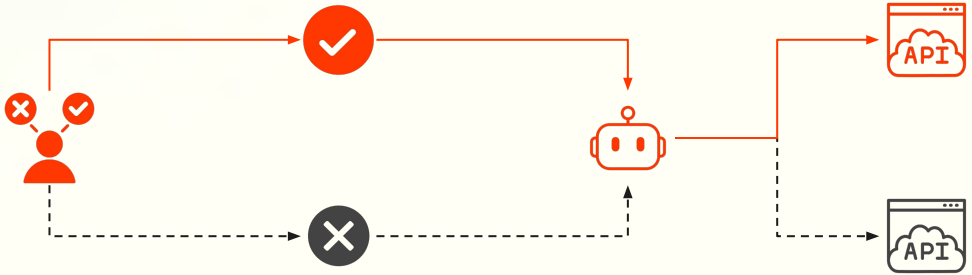
Utilizar de técnicas de prompt engineering para garantir que o prompt utilizado pelo modelo contenha as informações relevantes e complementares à base de conversas mencionada anteriormente;

Utilizar-se de técnicas de RLHF (reinforcement learning from human feedback) para treinar o modelo sempre que necessário, mantendo-o atualizado e com evoluções constantes.



DESAFIO 6 - INTEGRAÇÕES

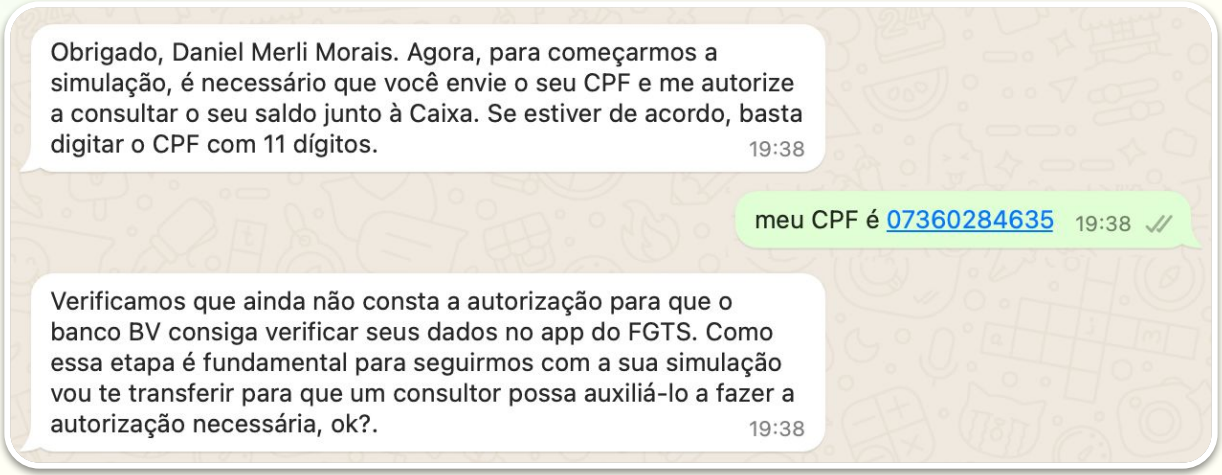
Como a AI participa do fluxo?



1. Decidindo qual API chamar com base na intenção do usuário
2. Fazendo o parsing das informações que se tornarão parâmetros da API
3. Fazendo o parsing do retorno e transformando-o em uma linguagem natural

DESAFIO 6 - INTEGRAÇÕES

Resultado final:



consulta?cpf=073.602.846-35

→

```
{
  code: 701
  description: ”CPF não autorizado”
}
```

DESAFIO 7 - EFICIÊNCIA EM CUSTO

Como tudo começou:

TOKEN EM LLM = CARO

AWS = CARO

FINOPS MINDSET = ZERO

DESAFIO 7 - EFICIÊNCIA EM CUSTO

Como estamos:

		Custo por 1M tokens	
Fornecedor	Modelo	Input	Output
OpenAI	GPT-3.5 Turbo (0125)	$0.50	$1.50
	GPT-4 Turbo (1106)	$10	$30
	GPT-4o	$5	$15
	GPT-4o-mini	$0.15	$0.60
Anthropic	Claude 3 Haiku	$0.25	$1.25
	Claude 3 Sonnet	$3	$15
Maritaca	Sábia 2 Small	R$1 (~$0.18)	R$3 (~$0.54)
	Sábia 3	R$10 (~$1.81)	R$10 (~$1.81)

AWS Bedrock

BR!

Preço e Qualidade

DESAFIO 7 - EFICIÊNCIA EM CUSTO

Outras considerações:

- Fine-tuning é bom, porém mais caro. Você realmente precisa?
- Reduza seu prompt (dica: LLMLingua)
- Use o modelo adequado para cada situação
- Registre seus custos ↓

id_llm_cost	fk_id_chat	feature_name	fk_id_llm_model	input_tokens	output_tokens	feat_request_id	price	created_at	updated_at
1025827	9253930	Analise de Emoção	6	345	15	NULL	0.000105	2024-08-28 00:43:59	2024-08-28 00:43:5
1025826	9570091	Analise de Emoção	6	333	15	NULL	0.000102	2024-08-28 00:43:53	2024-08-28 00:43:5
1025825	9253930	Analise de Emoção	6	346	17	NULL	0.00010775	2024-08-28 00:43:36	2024-08-28 00:43:3
1025824	9253930	Analise de Emoção	6	346	15	NULL	0.00010525	2024-08-28 00:43:20	2024-08-28 00:43:2
1025823	10130560	Analise de Emoção	6	328	17	NULL	0.00010325	2024-08-28 00:43:00	2024-08-28 00:43:0
1025822	10131179	API	1	126	16	NULL	0.000087	2024-08-28 00:42:58	2024-08-28 00:42:5
1025821	10131179	API	1	110	31	NULL	0.0001015	2024-08-28 00:42:56	2024-08-28 00:42:5
1025820	10131340	Analise de Emoção	6	327	17	NULL	0.000103	2024-08-28 00:42:26	2024-08-28 00:42:2
1025819	9823894	Analise de Emoção	6	332	18	NULL	0.0001055	2024-08-28 00:42:15	2024-08-28 00:42:1
1025818	10122989	Analise de Emoção	6	328	17	NULL	0.00010325	2024-08-28 00:42:13	2024-08-28 00:42:1

Dúvidas 🤔

Muito obrigado !!!