

# Inteligência Artificial Generativa: criando soluções reais com RAG

*Como ensinar uma IA a responder com os seus dados*

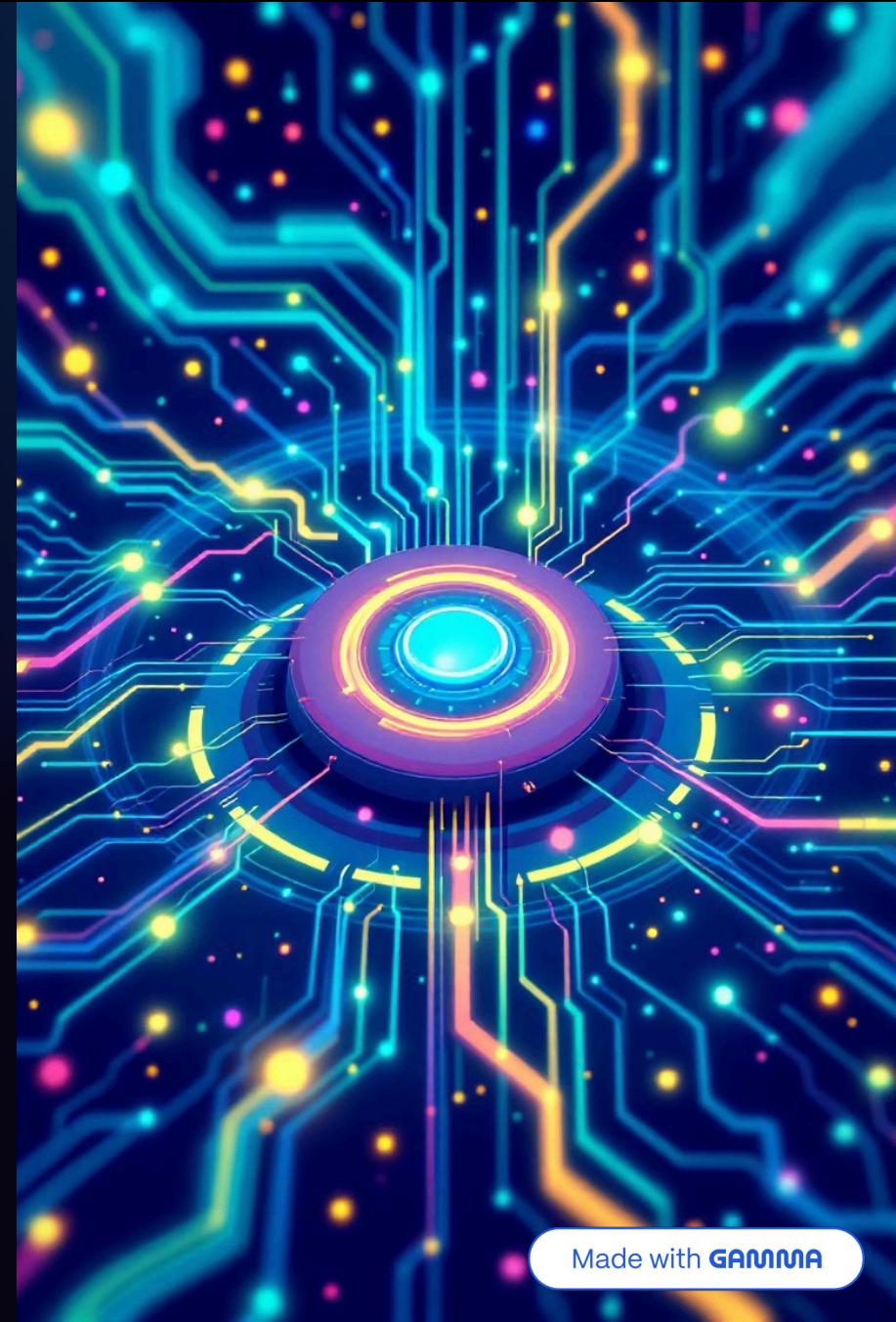


**José Luiz Orestes Junior**

Senior Engineering Manager

Certificado AWS | MBA em Projetos/Gestão de TI |

8+ anos liderando equipes globais | Entusiasta de IA,  
apaixonado por pessoas e tecnologia.



# O Novo Ciclo Tecnológico

## A evolução digital

Vivemos uma nova era de transformação: da Internet ao Mobile, e agora à IA Generativa. A diferença crucial? A IA não apenas analisa dados – ela **cria** conteúdo, escreve código e soluciona problemas complexos.

Essa revolução está transformando radicalmente como estudamos, trabalhamos e desenvolvemos software, criando oportunidades sem precedentes.

## Impacto Real

A IA generativa já revoluciona setores como educação, saúde, finanças e TI. Ferramentas como ChatGPT e copilotos de código mudaram o cotidiano de milhões de desenvolvedores.

**Pergunta para reflexão:** Quem aqui **NUNCA** usou IA generativa no dia a dia?



# O que são Large Language Models (LLMs)?

São modelos de Inteligência Artificial treinados com grandes volumes de texto – livros, artigos, código, conversas – para entender, gerar e processar linguagem humana com alta complexidade.

**Em outras palavras:** LLMs aprendem a reconhecer padrões, estruturas e significados na linguagem, permitindo que gerem texto coerente e relevante para diversas aplicações.

Exemplos: GPT · Claude · Gemini · LLaMA · Mistral



# GPT vs ChatGPT: Entendendo a Diferença

É como comparar um motor potente com um carro completo: ambos são importantes, mas servem a propósitos diferentes.



## GPT (Generative Pre-trained Transformer)

É o modelo de linguagem fundamental. GPTs são a 'inteligência bruta', o cérebro que aprende a linguagem através de um vasto treinamento de dados. Ele entende e gera texto de forma impressionante, mas por si só não tem memória de conversas passadas nem uma interface de usuário. É o motor.



## ChatGPT

É uma aplicação construída sobre um GPT (ou outro LLM). Ele adiciona uma camada de interface conversacional, memória de diálogo e ajustes finos para tornar a interação mais fluida e útil. É o carro completo, pronto para ser dirigido, com funcionalidades adicionais para o usuário final.

# Os Limites dos LLMs

LLMs são extremamente poderosos, mas enfrentam desafios críticos que podem limitar sua aplicação prática em ambientes corporativos.

## Memória Limitada

LLMs "esquecem" informações específicas e possuem conhecimento congelado no tempo de treinamento.

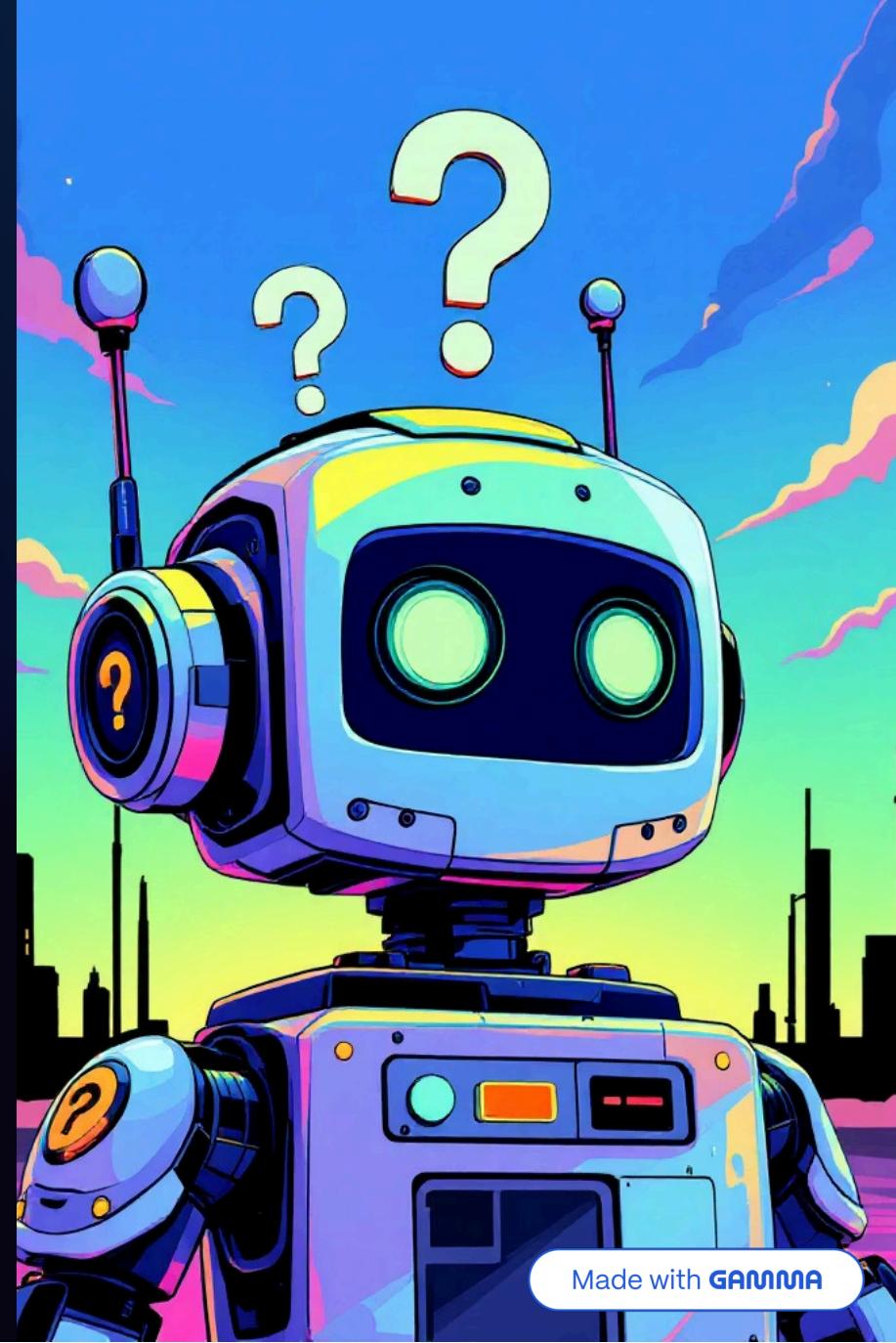
## Dados Privados

Não têm acesso a documentos internos, políticas corporativas ou informações proprietárias da sua empresa.

## Alucinações

Podem inventar respostas plausíveis mas incorretas quando não conhecem a informação solicitada.

- ❑ **Desafio real:** Como fazer a IA responder com base nos *seus próprios dados*? É aqui que entra o **RAG**.



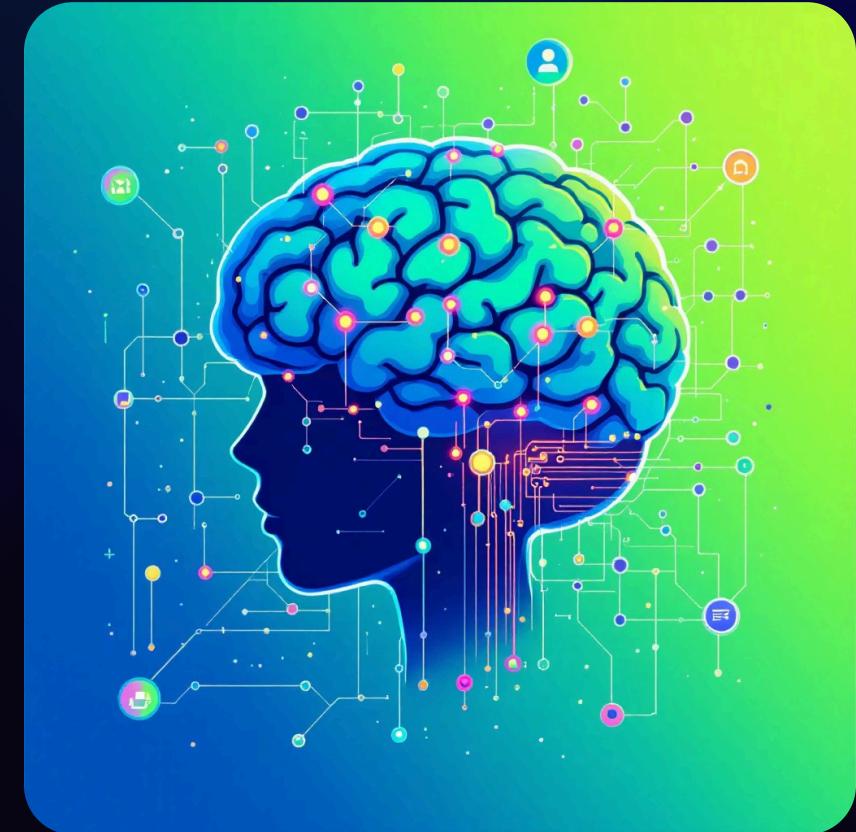
# RAG: Memória de Longo Prazo para IAs

## O que é:

Retrieval Augmented Generation (RAG) é uma arquitetura que permite à IA consultar bases de conhecimento externas ou internas antes de gerar respostas.

## Por que é importante:

Combina a criatividade do LLM com a precisão e atualização dos dados específicos. Permite respostas mais confiáveis e contextualizadas, mesmo para informações recentes ou detalhadas.



# VectorDB: A Tecnologia por Trás da Busca Inteligente

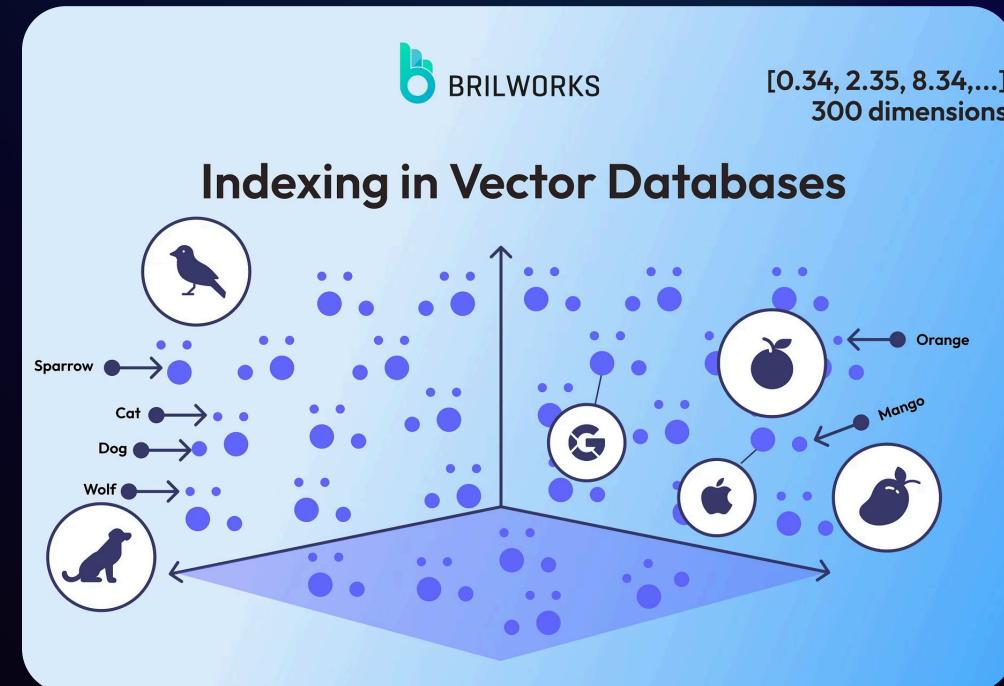
## Conceito Fundamental

Um Vector Database armazena **embeddings** – representações numéricas de texto que capturam significado semântico.

Embeddings próximos no espaço vetorial = significados semelhantes

## Diferencial

Enquanto um banco relacional busca **igualdades**, um Vector DB busca **proximidades de significado**.



# Como funciona a Arquitetura RAG?

A arquitetura Retrieval Augmented Generation (RAG) integra um Large Language Model (LLM) com uma base de conhecimento externa, permitindo que a IA recupere informações relevantes antes de gerar uma resposta. Este processo garante que as respostas sejam precisas, atualizadas e baseadas nos seus dados específicos.



## 1. Usuário

Faz uma pergunta ou consulta.

## 2. Embedding

Converte a pergunta em um vetor numérico (representação matemática).

## 3. Vector DB

Busca vetores semelhantes (trechos relevantes de seus documentos) na base de dados.



## 4. Contexto

Constrói o texto de apoio com as informações recuperadas para o modelo.



## 5. LLM

Gera a resposta usando o contexto fornecido, combinado com seu conhecimento geral.



## 6. Resultado

Apresenta uma resposta fundamentada e personalizada.

# Demo: Chat com Documentos



□ Próxima etapa: Live Coding - Demo ao vivo consultando documentos com RAG

# Implementação Prática



## Python

Linguagem de programação principal para desenvolvimento e orquestração.



## LangChain

Framework para construção de aplicações com LLMs, facilitando a integração dos componentes.



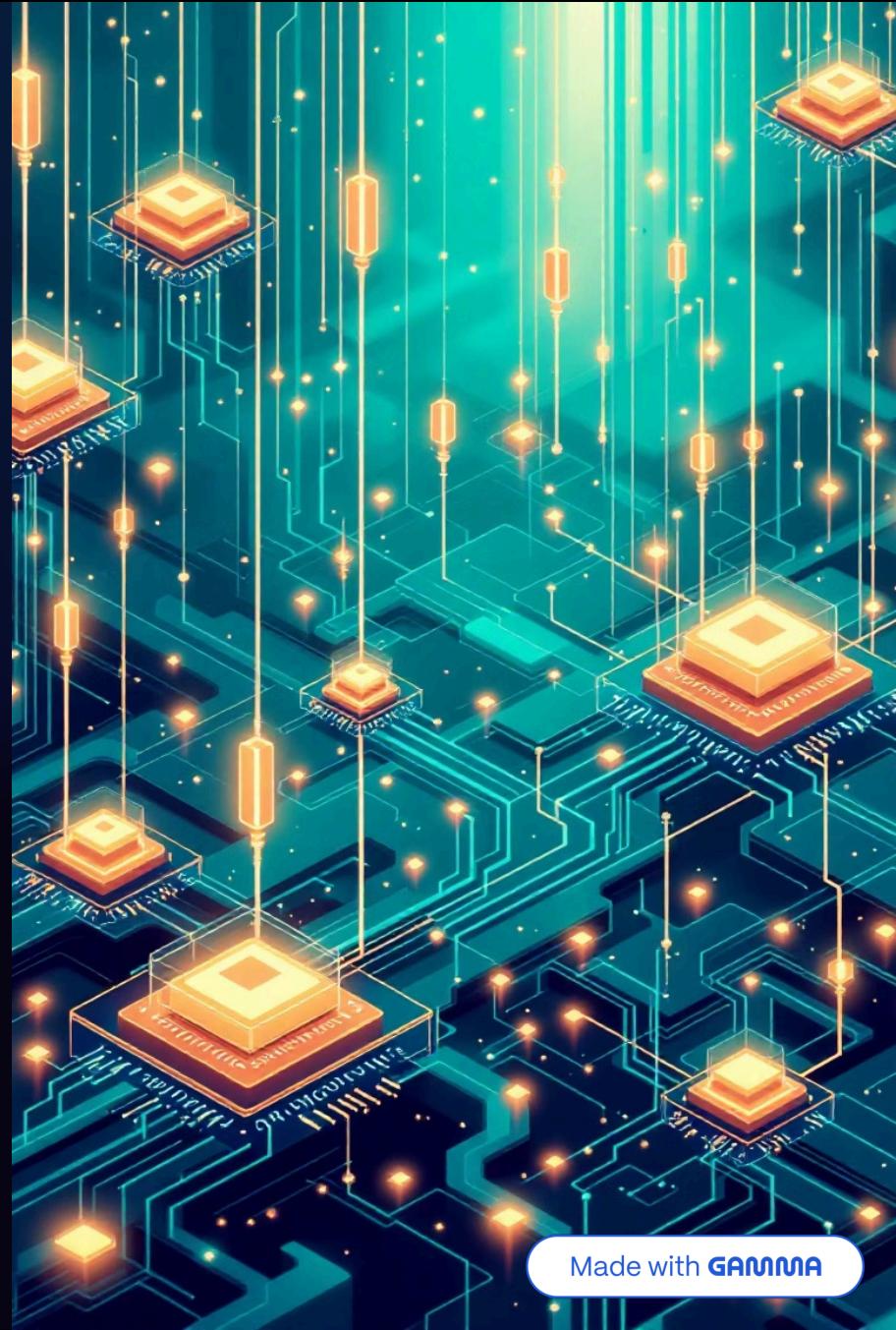
## Chroma

Bancos de dados vetoriais para armazenamento e busca de embeddings.



## OpenAI API

Interface para acessar modelos de linguagem da OpenAI, como GPT-3.5 e GPT-4.



Made with GAMMA

# Aplicações Reais de RAG

## Casos de Uso Transformadores



### Chatbots Corporativos

Assistentes virtuais que entendem políticas internas, processos e documentação técnica da empresa.



### Pesquisa Acadêmica

Ferramentas inteligentes para busca e análise de papers, teses e literatura científica.



### FAQ Dinâmico

Sistemas de perguntas frequentes que aprendem e se atualizam automaticamente.



### Suporte e Produtividade

Assistentes que aceleram resolução de tickets e aumentam eficiência operacional.

# Oportunidades de Carreira em IA Generativa

1

## Prompt Engineer

Especialista em criar e otimizar prompts para LLMs, buscando respostas precisas.

2

## Data Engineer for AI

Constrói e mantém pipelines de dados, incluindo Vector Databases, para alimentar modelos de IA.

3

## MLOps/AI Ops

Gerencia a implantação, monitoramento e ciclo de vida de modelos de IA em produção, visando escalabilidade.

4

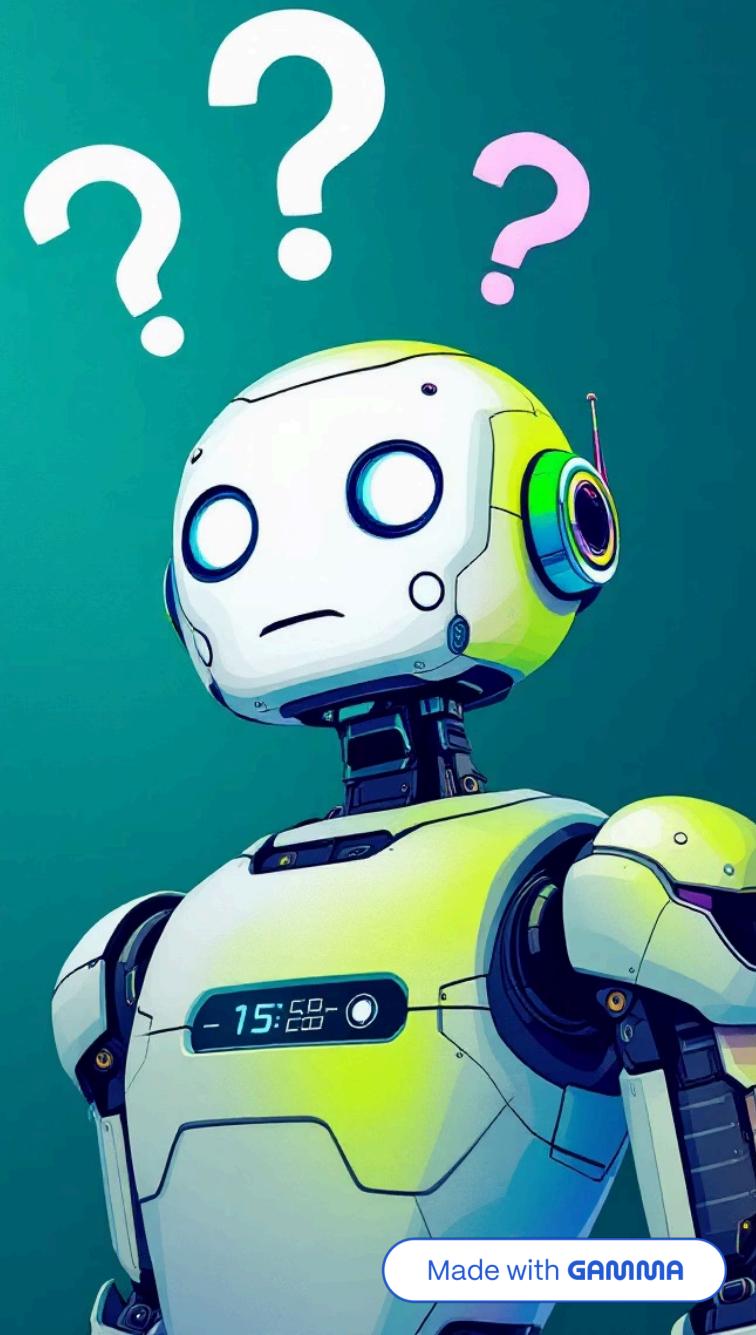
## Desenvolvedor GenAI

Cria e integra aplicações inovadoras usando APIs e frameworks de IA Generativa.



Made with GAMMA

# Dúvidas?



Made with GAMMA

O RAG é o que transforma  
um modelo genérico em  
uma inteligência que  
realmente entende o seu  
mundo.

---



## Contato

Conecte-se comigo no LinkedIn, escaneie o QR code ao lado para acessar meu perfil.

