



Do what matters

cogna
EDUCAÇÃO

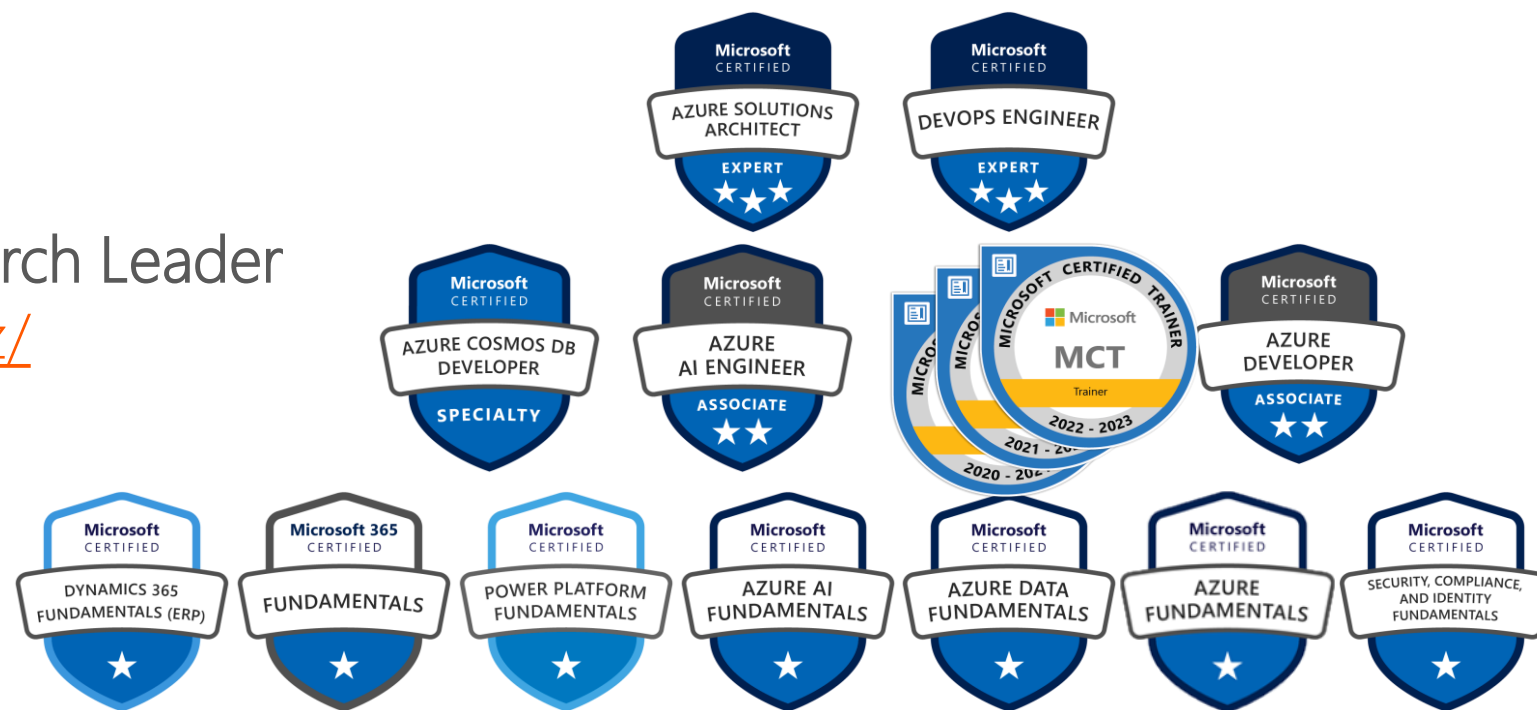
ChatGPT e OpenAI – Como integrar a tecnologia em suas aplicações

Diretor Senior – Enterprise Tech Arch Leader

<https://www.linkedin.com/in/aracz/>

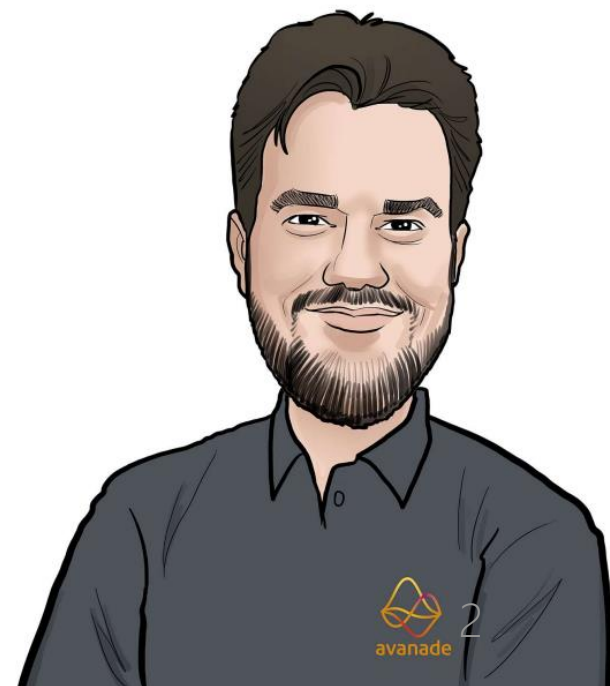
<https://github.com/andrerracz/>

- 23 anos de carreira
- + de 16 anos em Arquitetura
- 5 anos de Avanade



Projetos em grandes empresas dos segmentos:
Bancos / Seguradoras / Setor Público / Indústrias
/ Farmacêutico / Hospitais / Energia / Telecom

Tecnologias: Cloud Native, Java, .Net, NodeJS,
Kubernetes, Cloud, Devops, IA, IA Generativa



O que é IA generativa

Algoritmos de inteligência artificial que usam **conteúdo existente** como texto, arquivos de áudio ou imagens **para criar novos conteúdos plausíveis**.



Exemplos de tecnologias:

OpenAI

- ChatGPT
- GPT-3.5
- Codex
- Whisper
- Jukebox
- Dall-E 2

GitHub Co-Pilot

Google "BARD"

Stable diffusion

Midjourney

Meta "Galaxy"

Principais modelos/serviços

OpenAI:

GPT-3.5: Modelo de linguagem grande baseado em 175 bilhões de parâmetros

- Chat GPT: chatbot de uso geral
- Codex: Completar/gerar código
- Whisper: Conversão de fala em texto
- DALL-E: geração de imagens guiada pela entrada de texto
- JukeBox: geração musical
- Open AI/API: acesso aos principais recursos do Open AI

História do Microsoft Azure e OpenAI

Projeto Red Dog lançado como Windows Azure



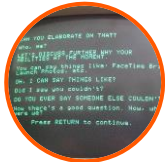
2010

Windows Azure renomeado como Microsoft Azure



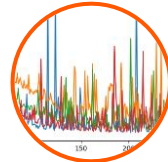
2014

Microsoft lança Serviços de ML do Azure



2015

Serviços Cognitivos do Azure – IA Massificada



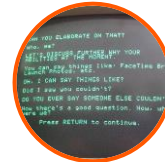
2018

Microsoft investe US\$ 1 bilhão na OpenAI



2019

Microsoft lança o Serviço Azure OpenAI



2021

A Microsoft investe US\$ 10 bilhões na Open AI. Prossegue com **Integração de OpenAI em todos os produtos Microsoft.**



2023

A OpenAI é fundada como um laboratório de pesquisa de IA de código aberto e sem fins lucrativos



OpenAI se abre para fins lucrativos, abrindo oportunidades de financiamento



OpenAI lança a IA pré-treinada generativa GPT-3



OpenAI Lança o DALL-E gerando imagens a partir de palavras



OpenAI incendeia o mundo com o ChatGPT



- 
[MSFT Bing AI Search](#)
- 
[MSFT Edge Web Co-pilot](#)
- 
[MSFT Bing AI Chat](#)
- 
[MSFT Viva Productivity](#)
- 
[Outlook Productivity](#)
- 
[Teams Premium Meeting Notes](#)
- 
[Azure OpenAI AlaaS](#)

Exemplo: Azure OpenAI Assistant vs um chatbot regular

O Azure OpenAI fornece até 99% de precisão de resposta em comparação com um criador de perguntas e respostas com cerca de 85% de precisão*

Chatbot regular:


Show me clinical data on how vitamin D does help with Covid19 prevention?

Just now

No answer found.

Just now

Type your message



Chatbot com Azure OpenAI:

Show me clinical data on how vitamin D does help with Covid19 prevention?

Data Source: [link](#)

In a double-blind trial conducted at 4 hospitals in Mexico, frontline health care workers were randomized to receive vitamin D3 4,000 IU or placebo for 30 days. Over one-third of the enrolled participants dropped out before study completion. Of the 192 participants who completed follow-up, 6.4% of participants in the vitamin D3 arm and 24.5% in the placebo arm acquired SARS-CoV-2 infection (relative risk 0.22; 95% CI, 0.08–0.59).

Just now

Type your message

O **Assistente baseado no Azure OpenAI** fornece uma resposta precisa (até 99%* de precisão de resposta) a partir da fonte de dados correta e lista essa fonte, enquanto o chatbot regular requer treinamento maciço para chegar perto desse nível.

covid19treatmentguidelines-389.pdf

1 / 1 | 90% + -

• There is insufficient evidence to recommend either for or against the use of vitamin D for the prevention or treatment of COVID-19.

Rationale

Vitamin D is critical for bone and mineral metabolism. Because the vitamin D receptor is expressed on immune cells such as B cells, T cells, and antigen-presenting cells, and because these cells can synthesize the active vitamin D metabolite, vitamin D also has the potential to modulate innate and adaptive immune responses.¹ It is postulated that these immunomodulatory effects of vitamin D could potentially protect against SARS-CoV-2 infection or decrease the severity of COVID-19.

Vitamin D deficiency (defined as a serum concentration of 25-hydroxyvitamin D <20 ng/mL) is common in the United States, particularly among persons of Hispanic ethnicity and Black race.² These groups are overrepresented among cases of COVID-19 in the United States.³ Vitamin D deficiency is also more common in older patients and patients with obesity and hypertension; these factors have been associated with worse outcomes in patients with COVID-19.⁴ High levels of vitamin D may cause hypercalcemia and nephrocalcinosis.⁵

Clinical Data on Vitamin D for Prevention

In a double-blind trial conducted at 4 hospitals in Mexico, frontline health care workers were randomized to receive vitamin D₃ 4,000 IU or placebo for 30 days.⁶ Participants were enrolled before COVID-19 vaccines became available. Over one-third of the enrolled participants dropped out before study completion. Of the 192 participants who completed follow-up, 6.4% of participants in the vitamin D₃ arm and 24.5% in the placebo arm acquired SARS-CoV-2 infection (relative risk 0.22; 95% CI, 0.08–0.59). At baseline, approximately 67% of participants had vitamin D deficiency, but this was not found to be an independent predictor of acquiring SARS-CoV-2 infection. The frequency of SARS-CoV-2 infection was very high in the placebo group, and it is unclear how these results translate to the use of vitamin D in vaccinated health care workers.

* Com base em um conjunto de perguntas de 104 perguntas e 5 documentos, os chatbots foram construídos sem otimizações manuais.

Original data source

Alguns cases de clientes no Brasil

Medicina Diagnóstica

iOCR para Pedidos médicos



Jornada

Agendamento de exames a partir do pedido médico

Objetivos

- **Redução do tempo** das análises para atendimento e retorno
- **Assertividade** na interpretação do pedido

Recursos IA Gen.

Contextualização

Classificação

Extração de Entidades

Hospital

Estruturação de prontuário



Jornada

Estruturação dos campos abertos de prontuários

Objetivos:

- **Viabilizar** a geração de insights
- **Redução no tempo** de resposta

Recursos IA Gen.

Sumarização

Classificação

Geração de Texto

Extração de Entidades

Hospital

Sugestão de Protocolo



Jornada

Sugestão de protocolo para atuação médica

Objetivos

- **Redução no tempo** de resposta
- **Assertividade** na interpretação dos dados

Recursos IA Gen.

Geração de Texto

Classificação

Busca Semântica

Alguns cases de clientes no Brasil

Demandas Judiciais Obrigações e Ofício



Jornada

Atendimento de ofício

Objetivos

- **Redução do tempo** das análises para atendimento e retorno
- **Assertividade** na interpretação do pedido

Recursos IA Gen.

Sumarização

Classificação

Extração de Entidades

Manifestações Sumarização e Causa Raiz (Ouvidoria)



Jornada

Atendimento 1ª instância (Fale Conosco e SAC)

Objetivos:

- **Redução de custos** operacionais
- **Redução no tempo** de resposta

Recursos IA Gen.

Sumarização

Classificação

Geração de Texto

Extração de Entidades

Venda Assistida (Consórcios)



Jornada

Simulação e venda de consórcio pelo gerente (Canal Link)

Objetivos

- Escalar as vendas digitais de consórcios, e proporcionar uma **experiência assistida e resolutive**

Recursos IA Gen.

Q&A

Classificação

Geração de Texto

Busca Semântica

Relacionamento com Clientes (AOC, Cartões PF e PJ)



Jornada

Localizar informações para Clientes PF, PJ & Cartões

Objetivos

- **Reduzir o acionamento** do time comercial
- **Diminuir o tempo** médio de retorno ao cliente

Recursos IA Gen.

Q&A

Geração de Texto

Busca Semântica

Causa Raiz Processos Judiciais (Jurídico)



Jornada

Emissão de pareceres consultivos

Objetivos

- **Simplificar** o esforço do Advogado em pesquisas com a qualidade dos pareceres emitidos
- **Diminuir o tempo** médio de respostas



Recursos IA Gen.

Q&A

Geração de Texto

Busca Semântica

IA responsável deve ser incorporada nos modelos

	Alucinações	Pensamento 'Black box'	Propenso a preconceitos	Preocupações com PI
 Desafios	Atualmente, os modelos LLM tendem a produzir respostas de som autoritário para perguntas, mesmo quando na verdade não sabe a resposta...	A saída dos modelos pode ser difícil de interpretar ou entender como o modelo produziu esse resultado.	Como qualquer solução de IA, as saídas são limitadas pela qualidade dos dados de origem. O viés dos dados de origem inseridos por humanos será extrapolado nas saídas	Dependendo de como os modelos são utilizados, ainda não há resultados claros de proteção da PI e da exposição ao uso acidental de outras PI utilizando saídas do modelo (especialmente em casos de uso de codificação).
 Desenvolvimento abordagens para mitigação	Cada aplicação de um LLM precisa de um ajuste cuidadoso com incorporações ou prompts para ajustar as respostas para casos de uso específicos. É mais provável que tenhamos LLMs especializados no futuro próximo do que inteligências gerais que resolvem todos os problemas.	O uso responsável e inteligente desses modelos será uma nova habilidade para o mundo. Levar o modelo a pensar "passo a passo" para mostrar seu funcionamento e ter controles apropriados em todas as etapas é crucial.	O uso responsável e inteligente desses modelos será uma nova habilidade para o mundo. Estamos colaborando com a Microsoft para um caso de uso de avaliação de impacto para LLMs - testando a estrutura de IA responsável da Microsoft	O envolvimento ativo no campo da IA generativa será crucial à medida que o cenário jurídico global se forma em torno do que é considerado aceitável e não aceitável do uso desses modelos.

Técnicas para interagir com LLMs

Three Patterns for Engineering and Integration with LLMs

Prompt Engineering

Dado um "prompt", complete o texto.. O prompt pode incluir instruções ou context para ajudar a responder uma pergunta.

Fine-Tuning

Use seus dados para ajustar os pesos de um modelo de linguagem pré-treinado, permitindo que o modelo armazene e "raciocine" sobre o seu texto. Também podem ser utilizados para adaptar o modelo para novas tarefas.

Embeddings

Usando um modelo base ou customizado, gere vetores que representam o estado interno do modelo baseado no seu texto. Permite que sejam feitas buscas por similaridade semântica.

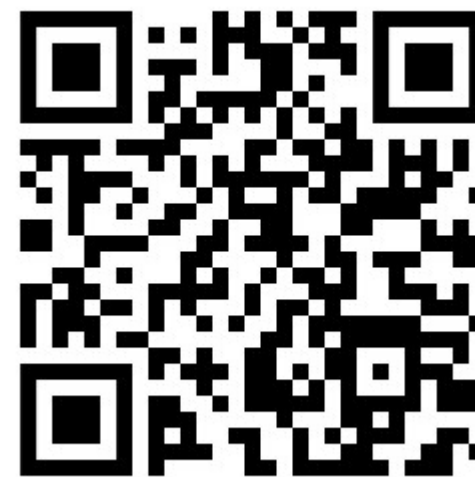


Do what matters

Chamando OpenAI a partir de código

Passos para rodar

1. Clonar o repositório: <https://github.com/andreracz/AzureOpenAITutorial>
2. Criar um arquivo .env dentro do diretório raiz, com o conteúdo:
 - › `OPENAI_API_BASE=https://your_url.openai.azure.com/`
 - › `OPENAI_API_KEY=KEY`
 - › `DEPLOYMENT_NAME=text-davinci-003`
3. Os exemplos estão numerados na sequência de execução.





Do what matters

Prompt Engineering

Prompt Engineering

Prompt Engineering são técnicas que podemos utilizar para obter um resultado melhor dos modelos de linguagem.

Técnicas:

ZERO SHOT

Traduza do Português para inglês:

Como você vai? ->

How are you going?

ONE SHOT

Traduza do Português para inglês:

Oi, tudo bem? ->Hi, how are you?
Como você vai?->

How are you doing?

FEW SHOT

Traduza do Português para inglês:

Oi, tudo bem? ->Hi, how are you?
Me chamo André -> My name is André
Como você vai? ->

How are you doing?

Algumas capacidades dos modelos

TRADUÇÃO

Traduza a seguinte frase para
1. Inglês, 2. Espanhol, 3.
Japonês: Por favor, gostaria de
uma Água

1.

Please, I would like some water.
2. Por favor, me gustaría algo de
agua.
3. お願いします、水が欲しいで
す。

GERAÇÃO DE IDÉIAS

Gere idéias de posts de blog
sobre Arquitetura de sistemas
para profissionais senior:

1. Como usar a arquitetura de sistemas para otimizar o desempenho de sistemas de computação em nuvem
2. O que os profissionais senior precisam saber sobre arquitetura de sistemas
3. Como implementar uma arquitetura de sistemas escalável
4. Os benefícios da arquitetura de sistemas para a segurança de dados
- ...

CLASSIFICAÇÃO

Decida se o Tweet abaixo é
positivo, negativo ou neutro.

Tweet: Estou adorando a nova
temporada do Mandalorian!
Sentimento:

Positivo

Decida se o Tweet abaixo é
positivo, negativo ou neutro.

Tweet: Estou adorando a nova
temporada do Mandalorian,
mas o último episódio foi bem
ruim!

Neutro

Algumas capacidades dos modelos

CLASSIFICAÇÃO E DETECÇÃO DE INTENÇÃO

Para o texto abaixo, dê duas classificações, das seguintes categorias:

- Fila de Atendimento: "Office", "E-mail", "Active Directory", "Sharepoint"
- Severidade: "Alta", "Média", "Baixa"

Título: Todos os usuários sem acesso

Estamos com TODOS os usuários da empresa sem acesso ao e-mail, as senhas não estão funcionando. Precisamos de ajuda imediata!

Resposta:

Fila de Atendimento: "Active Directory", "E-mail"

Severidade: "Alta"

* Texto completo extraído do link: <https://kubernetes.io/pt-br/docs/concepts/overview/what-is-kubernetes/>

Algumas capacidades dos modelos

SUMARIZAÇÃO

Resuma o texto abaixo e diga quais os principais pontos sobre ele.

O que é Kubernetes?

Kubernetes é uma plataforma de código aberto, portátil e extensiva para o gerenciamento de cargas de trabalho e serviços distribuídos em contêineres, que facilita tanto a configuração declarativa quanto a...

Kubernetes é uma plataforma de código aberto, portátil e extensiva para o gerenciamento de cargas de trabalho e serviços distribuídos em contêineres. Ele foi criado pelo Google em 2014, e seu nome tem origem no Grego, significando timoneiro ou piloto. O Kubernetes oferece aos usuários descoberta de serviço e balanceamento de carga, orquestração de armazenamento, lançamentos e reversões automatizadas, empacotamento binário automático, autocorreção, gerenciamento de configuração e de segredos. Ele não é um sistema PaaS (plataforma como serviço) tradicional e completo, mas fornece os blocos de construção para a construção de plataformas de desenvolvimento.

* Texto completo extraído do link: <https://kubernetes.io/pt-br/docs/concepts/overview/what-is-kubernetes/>

Algumas capacidades dos modelos

SUMARIZAÇÃO

Resuma o texto abaixo e explique para uma criança de 5 anos.

O que é Kubernetes?

Kubernetes é uma plataforma de código aberto, portátil e extensiva para o gerenciamento de cargas de trabalho e serviços distribuídos em contêineres, que facilita tanto a configuração declarativa quanto a...

Kubernetes é uma plataforma que nos ajuda a gerenciar nossas aplicações, como jogos, vídeos e outras coisas que usamos no computador. Ele faz isso usando contêineres, que são como caixinhas que armazenam nossas aplicações e nos ajudam a mantê-las seguras. O Kubernetes também nos ajuda a gerenciar nossas aplicações, para que elas funcionem bem e não sejam interrompidas.

* Texto completo extraído do link: <https://kubernetes.io/pt-br/docs/concepts/overview/what-is-kubernetes/>

Algumas capacidades dos modelos

SUMARIZAÇÃO

Resuma o texto abaixo e explique para uma criança de 5 anos.

O que é Kubernetes?

Kubernetes é uma plataforma de código aberto, portátil e extensiva para o gerenciamento de cargas de trabalho e serviços distribuídos em contêineres, que facilita tanto a configuração declarativa quanto a...

Kubernetes é uma plataforma que nos ajuda a gerenciar nossas aplicações, como jogos, vídeos e outras coisas que usamos no computador. Ele faz isso usando contêineres, que são como caixinhas que armazenam nossas aplicações e nos ajudam a mantê-las seguras. O Kubernetes também nos ajuda a gerenciar nossas aplicações, para que elas funcionem bem e não sejam interrompidas.

* Texto completo extraído do link: <https://kubernetes.io/pt-br/docs/concepts/overview/what-is-kubernetes/>

Algumas capacidades dos modelos

EXTRAÇÃO DE ENTIDADES

Extraia o nome da pessoa, nome da empresa, localidade, telefone e assunto do texto abaixo e formate como JSON.

Olá, meu nome é João da Silva, estou ligando da Mineiradora Contoso, em Vitória. Estou ligando pois meu e-mail não está funcionando. Por favor me ligue no telefone: 011 9555-1234.

```
```json  
{
```

```
 "nome": "João da Silva",
 "empresa": "Mineiradora Contoso",
 "localidade": "Vitória",
 "telefone": "011 9555-1234",
 "assunto": "E-mail não funcionando"
```

```
}
```
```



CoT – Chain of Thought

Induz o modelo a uma resposta certa ao pensar passo a passo.

SEM CoT

Q: A cafeteria tinha 23 maçãs.
se eles usaram 20 para
preparar o almoço e
compraram mais 6, quantas
maçãs eles tem?

R: 29 maçãs.

COM CoT

Q: A cafeteria tinha 23 maçãs.
se eles usaram 20 para
preparar o almoço e
compraram mais 6, quantas
maçãs eles tem?
De a resposta passo a passo,
fazendo a conta a cada passo.

1. Primeiro, temos 23 maçãs.
2. Depois, eles usaram 20 maçãs para preparar o almoço.
3. Então, restam 3 maçãs.
4. Em seguida, eles compraram mais 6 maçãs.
5. Finalmente, eles têm 9 maçãs.

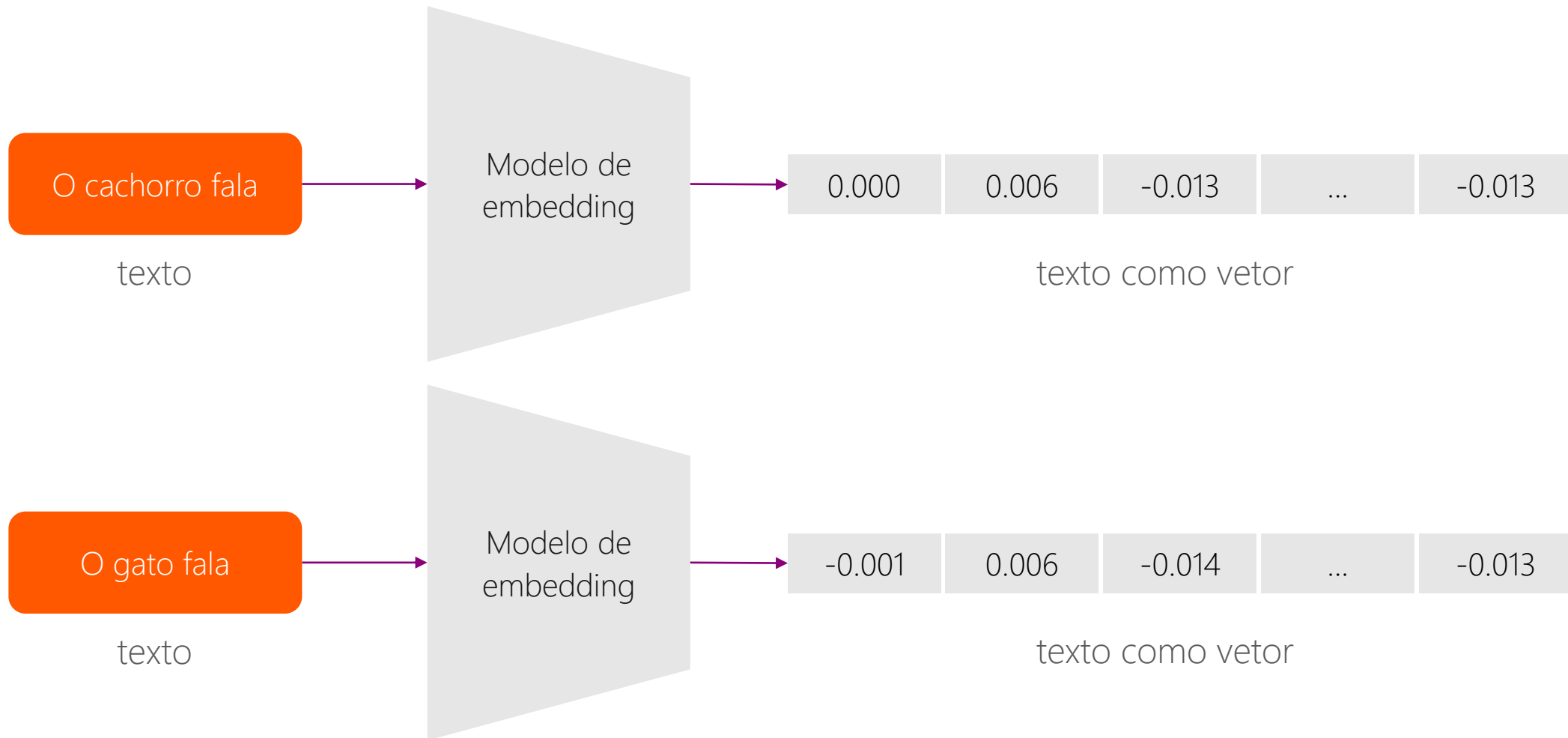


Do what matters

Embeddings

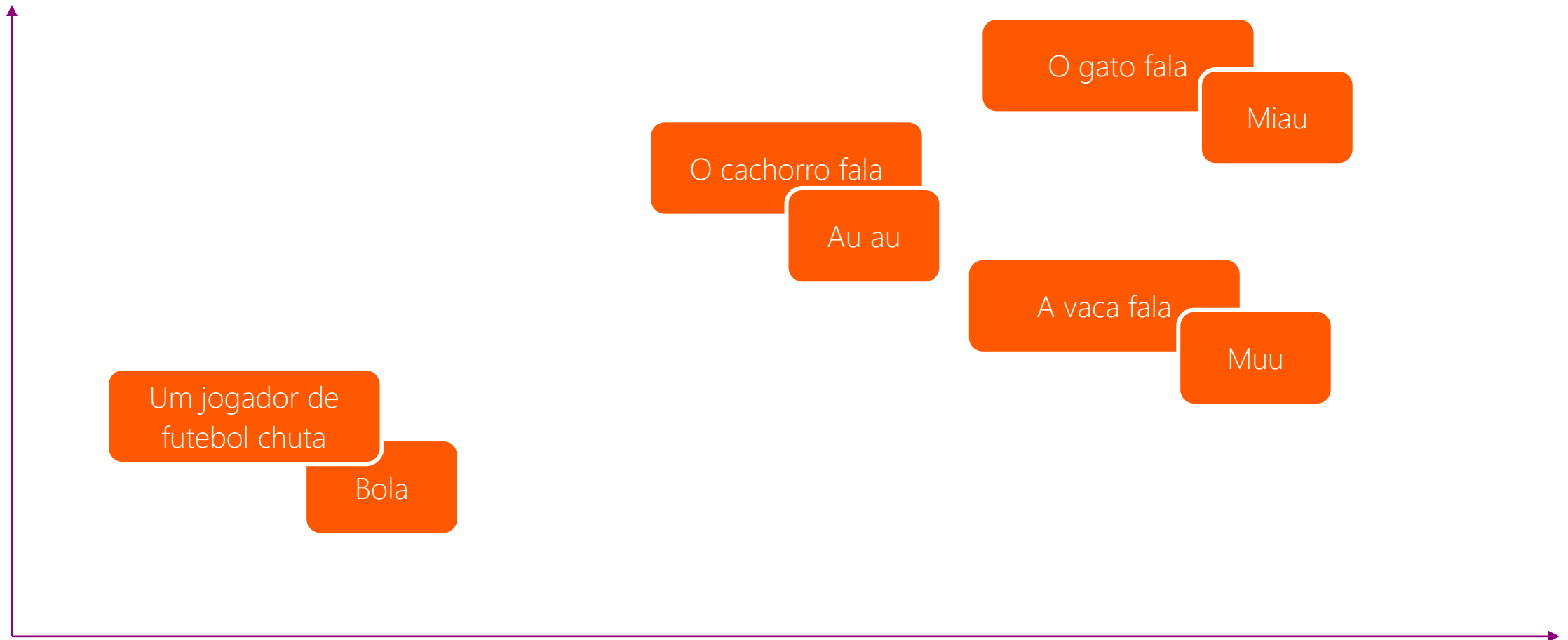
Embeddings

Entra texto, sai um vetor com a sua representação



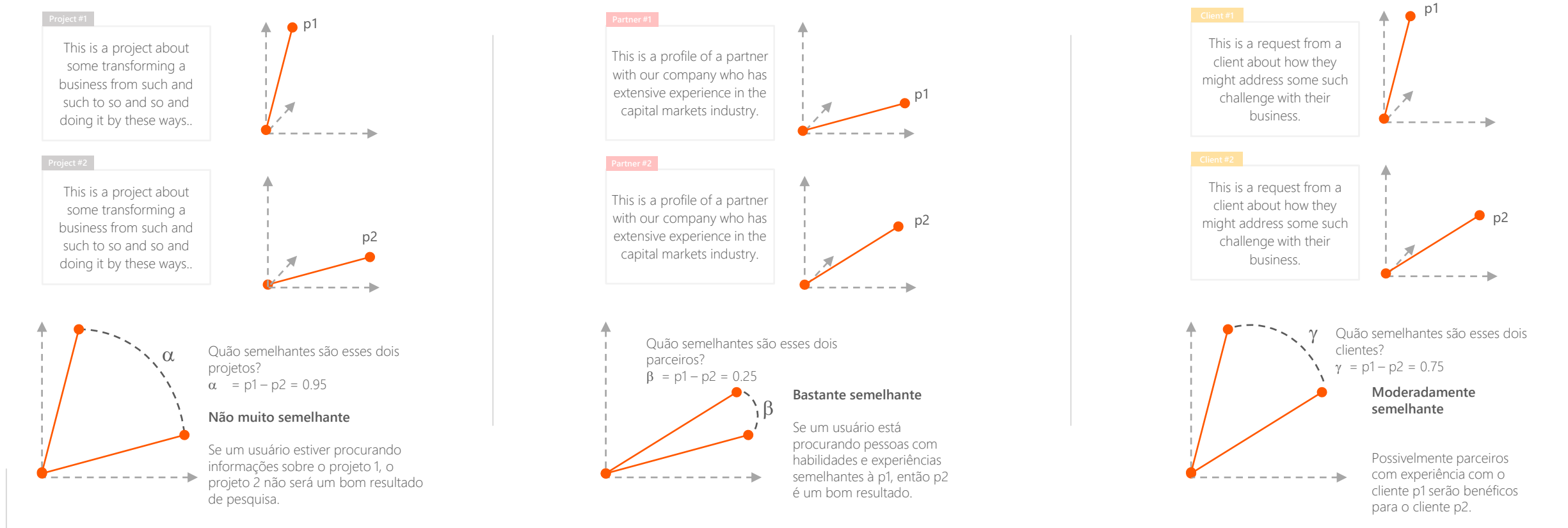
Embeddings

Resultado do vetor tem proximidade semantica com outros textos



Embeddings

Resultado do vetor tem proximidade semantica com outros textos



Embeddings permitem generalizar a pesquisa para além do texto, aproveitando a representação interna do texto dentro do modelo de linguagem. Você também pode usar os Embeddings para criar modelos específicos de domínio que são muito menores e generalizam mais rapidamente.



Dúvidas ou Comentários?