

# Explorando LMS AI

Benevid F. Silva



# CONTEÚDO



1

Visão geral LLMs

2

Aplicações

3

Termos comuns

4

Assistentes de IA

5

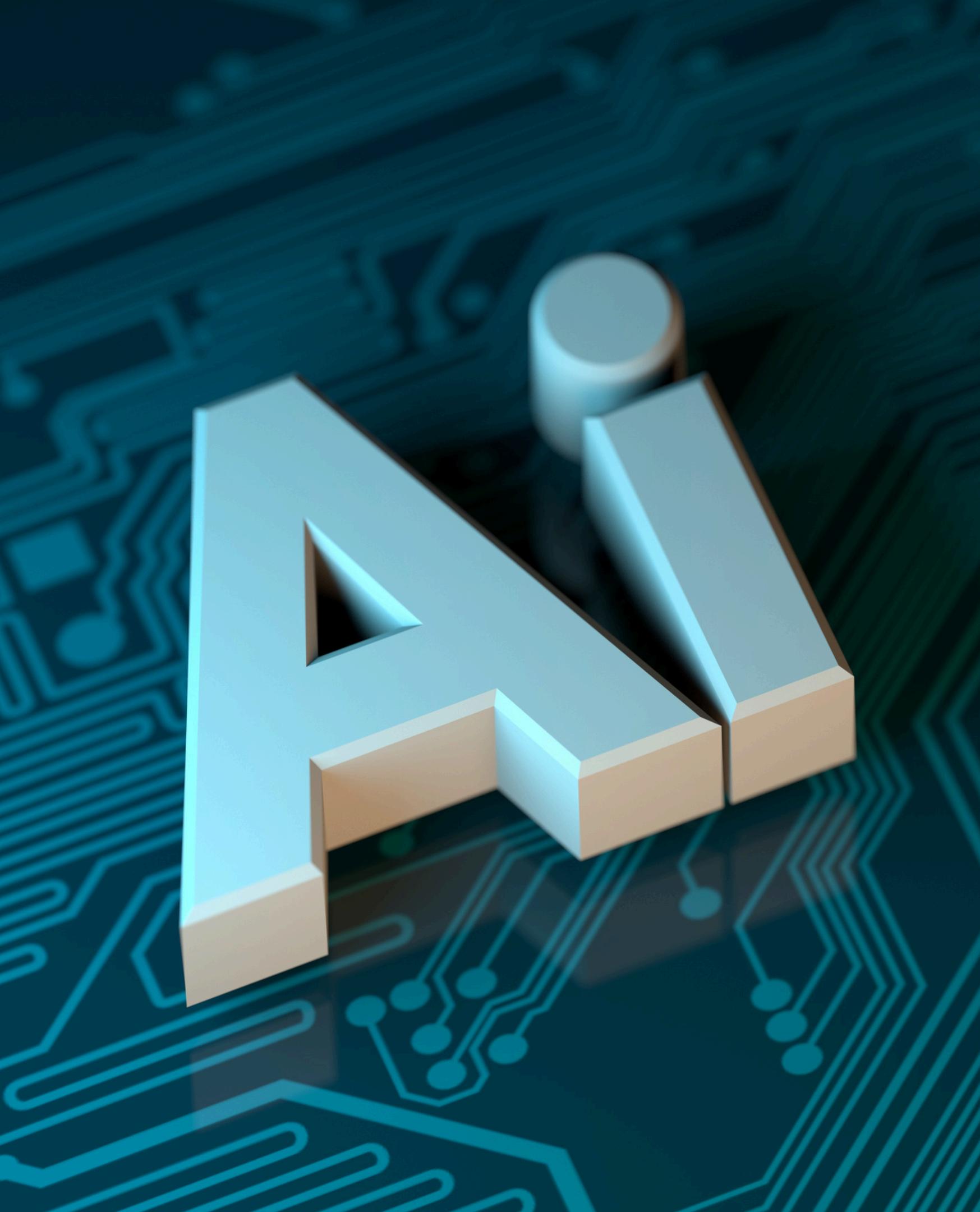
Prompt

6

Outros

# O QUE SÃO LLMS?

LLM (Large Language Models) são modelos de inteligência artificial (IA) que utilizam técnicas de Machine Learning (ML) para entender e gerar linguagem humana, como textos e imagens.





# COMO FUNCIONAM?

os LLMs utilizam um método chamado de **aprendizado não-supervisionado**. Nesse processo, um modelo de machine learning é alimentado com conjuntos de dados (centenas de **bilhões** de palavras e frases) os quais são estudados e aprendidos com base em exemplo.

Essa fase de aprendizado não supervisionado anterior ao treinamento é fundamental para o desenvolvimento de LLMs como o GPT (Generative Pre-Trained Transformer) e BERT (Bidirectional Encoder Representations from Transformers).

## TRANSFORMERS

- uma arquitetura de rede neural introduzida em 2017 por Vaswani et al. no artigo "Attention is All You Need". Essa arquitetura revolucionou o campo do Natural Language Processing (NLP) e é a base de muitos LLMs (Large Language Models).

## COMO FUNCIONAM OS TRANSFORMERS

**Generative** (deep learning) **models** for understanding and generating text, images and many other types of data.

**Transformers** analyze chunks of data, called "tokens" and **learn to predict the next token** in a sequence, **based on previous and**, if available, **following tokens**.

The **auto-regressive** concept means that the output of the model, such as the prediction of a word in a sentence, is influenced by the previous words it has generated.

## RELAÇÃO COM LLMS

- GPT (Generative Pre-trained Transformer), Gemini, entre outros, são construídos sobre a arquitetura Transformer

Music—MusicLM (Google) and Jukebox (OpenAI) generate music from text.

Image—Imagen (Google) and DALL.E (OpenAI) generate novel images from text.

Texte—OpenAI's GPT has become widely known, but other players have similar technology (including Google, Meta, Anthropic and others).

Others—Recommender (movies, books, flight destinations), drug discovery...



# RECURSOS

Como estão sempre calculando probabilidades para encontrar conexões, os LLMs exigem um volume significativo de recursos computacionais.

Uma maneira de obter a capacidade computacional necessária é por meio das unidades de processamento gráfico (GPUs).



Placa de vídeo NVIDIA Tesla H100 80GB oficial GPU computação não SXM-mostrar tít

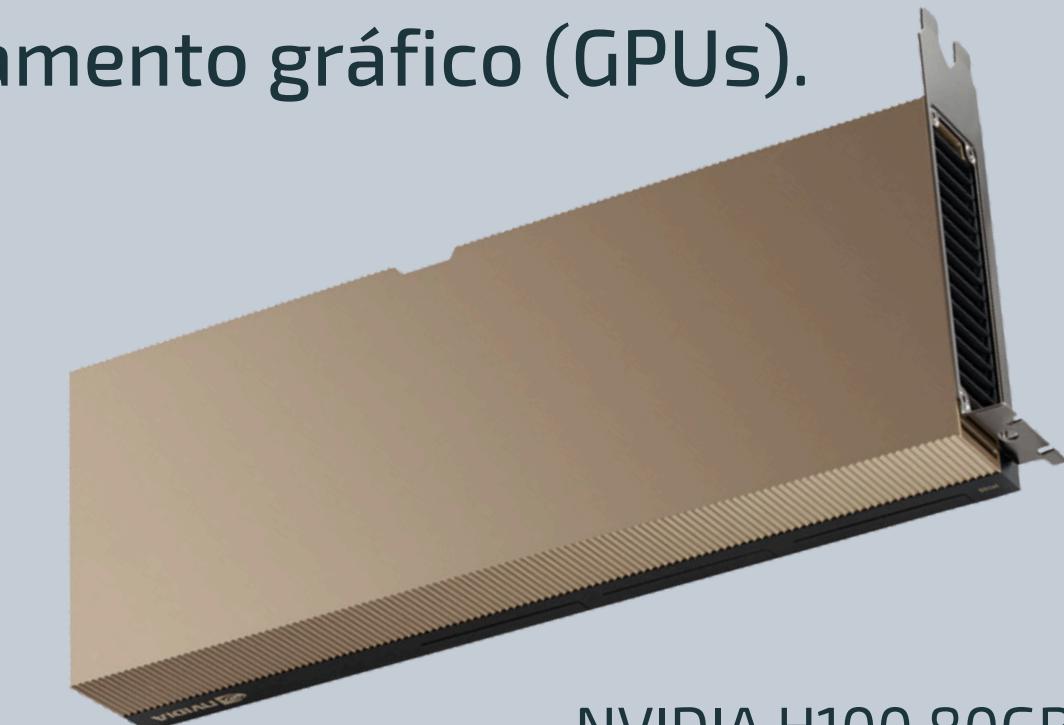


Lucky-PLC Store (930)

93.9% de avaliações positivas · Outros itens do vendedor

US \$45 000,00

Aproximadamente R\$ 244 863,31



NVIDIA H100 80GB



# TERMOS COMUNS

## FINE-TUNING

Processo de ajuste de um modelo pré-treinado em um novo conjunto de dados específico para uma tarefa particular, melhorando a performance do modelo naquela tarefa.

## ZERO-SHOT

Habilidade de um modelo realizar tarefas que não foram explicitamente treinadas, baseando-se apenas na compreensão pré-treinamento.

## TOKENIZATION

Processo de dividir texto em unidades menores, como palavras ou subpalavras, chamadas tokens, que são usados como entrada para LLMs.

## INFERENCE:

Processo de usar um modelo treinado para fazer previsões ou gerar texto com base em novas entradas

## EMBEDDING

Representação de palavras ou frases em um espaço vetorial contínuo, onde palavras com significados semelhantes estão mais próximasumas das outras

## RAG (RETRIEVAL-AUGMENTED GENERATION)

Técnica que combina a geração de texto por modelos de linguagem com a recuperação de informações de uma base de dados externa.



# TERMOS COMUNS

## PROMPT

Entrada ou instrução dada a um LLM para gerar uma resposta. Um prompt pode ser uma pergunta, uma frase inicial, ou qualquer tipo de comando que guia o modelo na geração de texto relevante.

## PROMPT ENGINEERING

Processo de formular e refinar prompts para otimizar a saída de um LLM.

## API

Interface que permite a integração de LLMs em outras aplicações.

## ALUCINAÇÃO

Um fenômeno onde o LLM gera informações incorretas ou irrelevantes.

## KNOWLEDGE BASE

Refere-se a sistemas que utilizam LLMs para criar, acessar ou expandir bases de conhecimento

## SENTIMENT ANALYSIS

Uma aplicação comum dos LLMs, onde eles classificam textos como positivos, negativos ou neutros, ajudando empresas a compreender as emoções dos clientes em feedbacks ou redes sociais

# LLMS MAIS COMUNS

## GPT-4 (GENERATIVE PRE-TRAINED TRANSFORMER 4)

Desenvolvido pela OpenAI, é a versão mais recente da série GPT, conhecida por sua capacidade de gerar texto de alta qualidade em uma ampla variedade de contextos

## GEMINI

Desenvolvido pelo Google DeepMind, Gemini é um LLM que integra as mais recentes pesquisas em IA para oferecer capacidades avançadas de geração e compreensão de texto.



A Hugging Face, uma das principais plataformas para modelos de IA, hospeda mais de **320.000** modelos diferentes

## CLAUDE

Criado pela Anthropic, Claude é um modelo de linguagem focado em segurança e alinhamento ético, projetado para minimizar vieses e maximizar a utilidade em tarefas de NLP.

## LLAMA (LARGE LANGUAGE MODEL META AI)

Desenvolvido pela Meta (anteriormente Facebook), o LLaMA é projetado para ser eficiente em termos de recursos, oferecendo desempenho robusto em várias tarefas de NLP com menos requisitos de computação

# 2022: ChatGPT

“**ChatGPT**, the popular chatbot from OpenAI, is estimated to have reached 100 million monthly active users in January, just two months after launch, making it the **fastest-growing consumer application in history**”

Reuters, Feb 1, 2023  
<https://reut.rs/3yQNIgo>

## PONTO DE MUDANÇA

## ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users

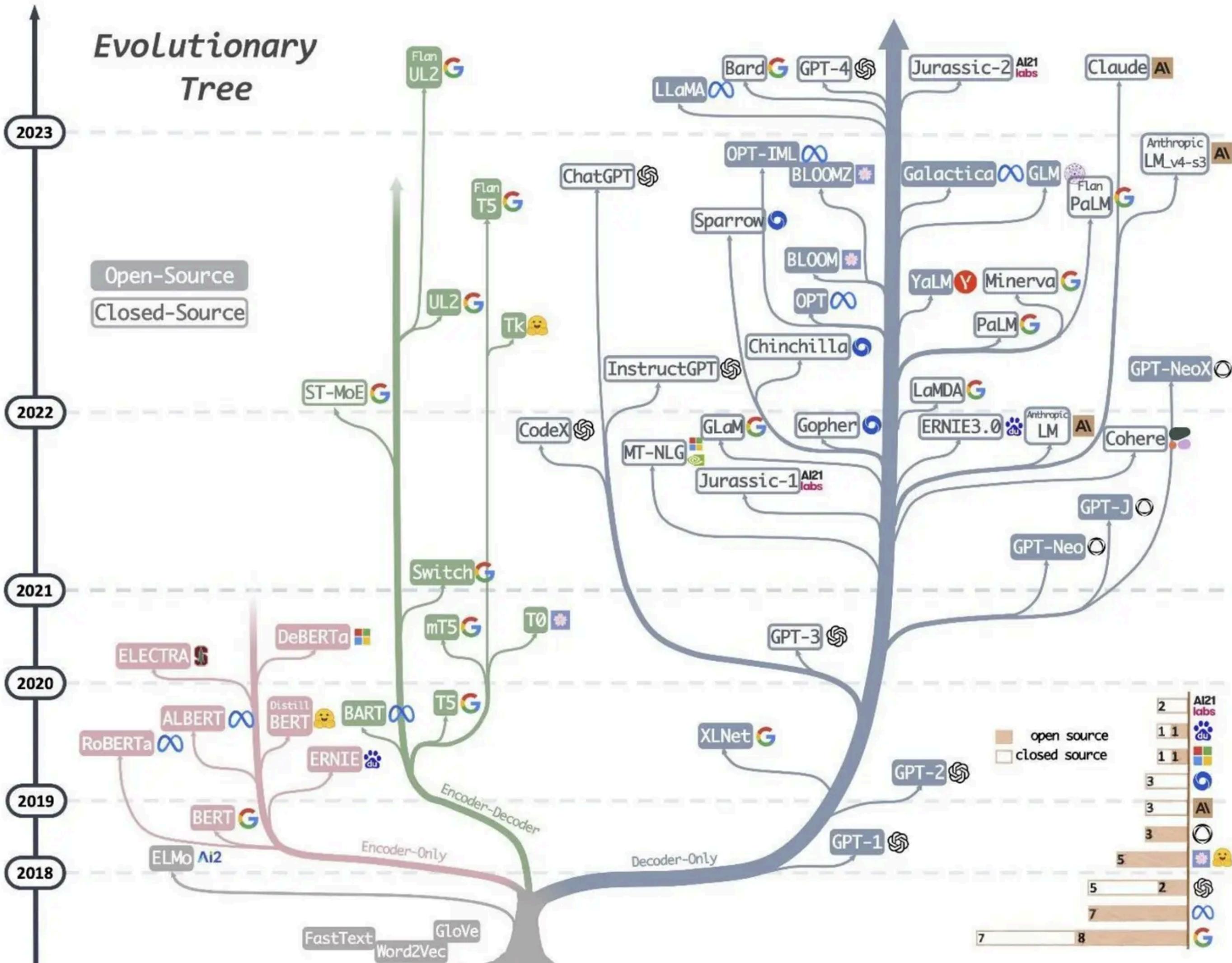


\* one million backers   \*\* one million nights booked   \*\*\* one million downloads  
Source: Company announcements via Business Insider/LinkedIn



statista

# EvoLutionary Tree

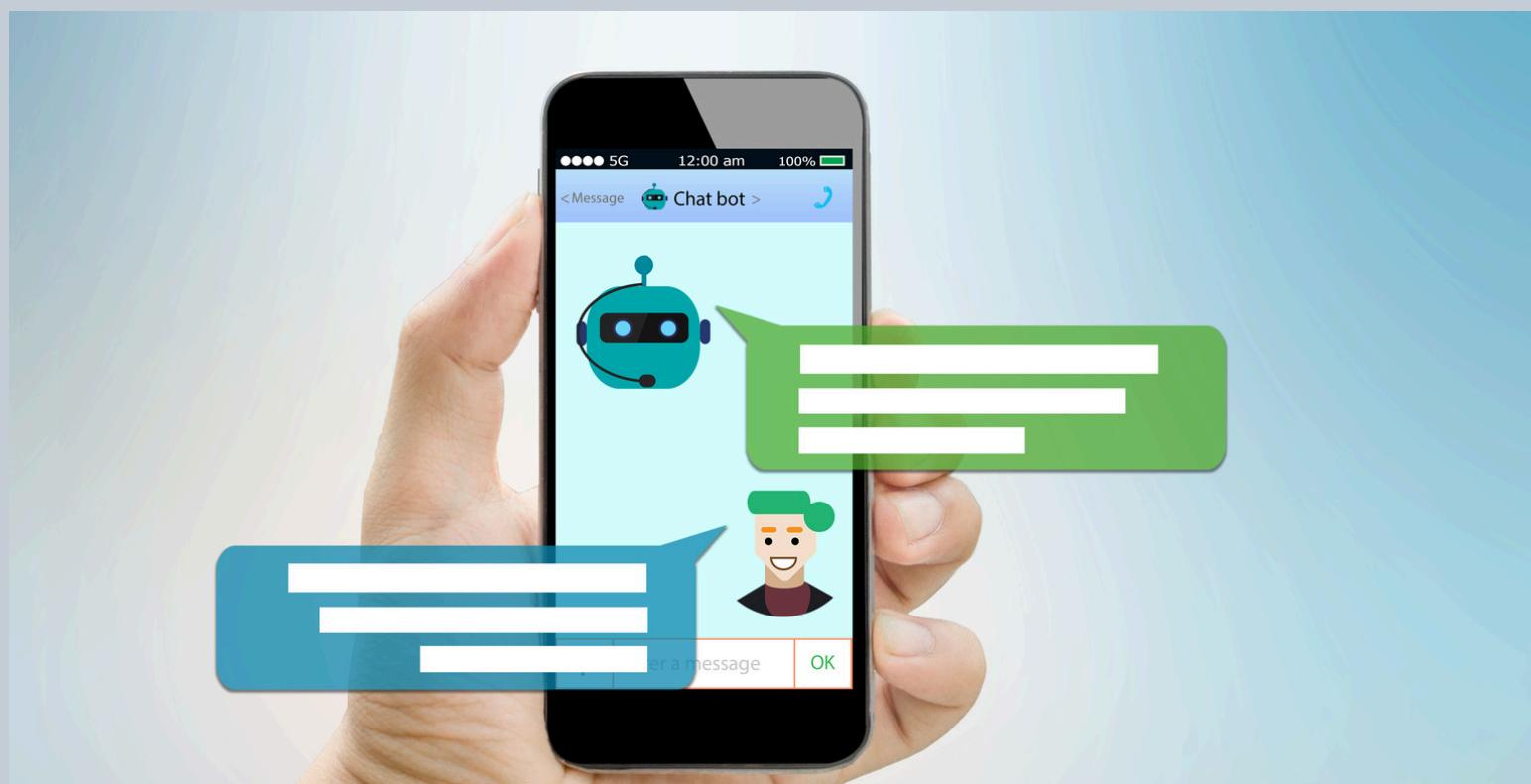


Source:  
[github.com/Mooler0410/LLMsPracticalGuide](https://github.com/Mooler0410/LLMsPracticalGuide)

# APLICAÇÕES

## GENERALISTAS

são modelos de linguagem de grande porte (LLMs, na sigla em inglês) treinados em vastas quantidades de dados textuais e de código, permitindo que eles gerem texto, traduzam idiomas, escrevam diferentes tipos de conteúdo criativo e respondam a suas perguntas de forma informativa.

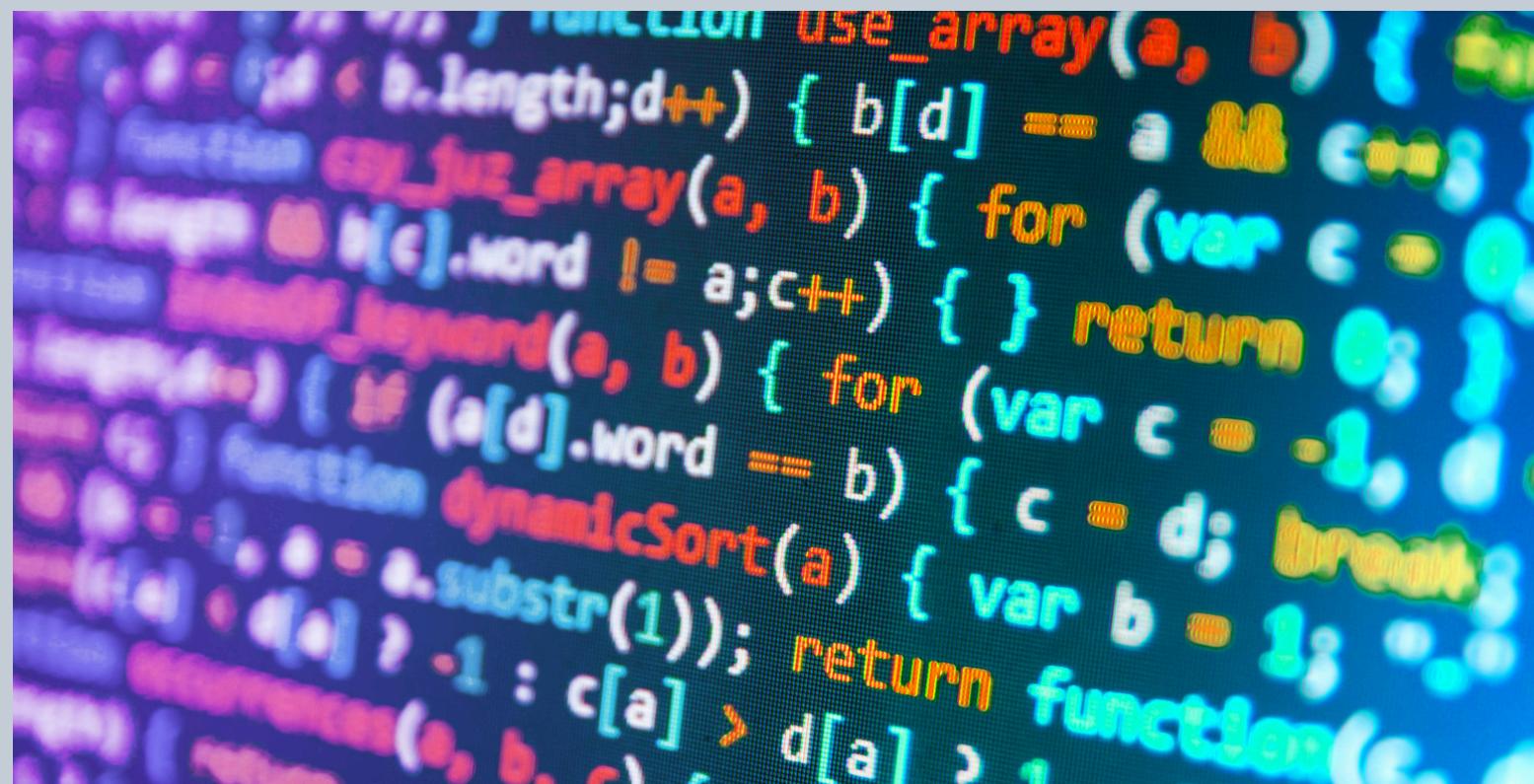


- **Geração de conteúdo:** Criação de artigos, posts de blog, scripts, código, etc.
- **Tradução automática:** Tradução de textos entre diferentes idiomas.
- **Respostas a perguntas:** Fornecimento de respostas concisas e informativas a perguntas complexas.
- **Chatbots e assistentes virtuais:** Criação de interfaces de conversação mais naturais e intuitivas.
- **Resumo de textos:** Geração de resumos concisos de textos longos.

# APLICAÇÕES

## NÃO GENERALISTAS

ao contrário dos modelos generalistas que são treinados em vastas quantidades de dados para realizar uma ampla gama de tarefas, são especializados em tarefas específicas. Eles são treinados em conjuntos de dados muito mais focados, permitindo que eles realizem essas tarefas com maior precisão e eficiência.



- **Modelos de geração de código:** Modelos que geram código em linguagens de programação específicas, como Python ou Java.
- **Modelos de análise de sentimentos em domínios específicos:** Modelos treinados para analisar sentimentos em textos relacionados a um determinado domínio, como reviews de produtos ou posts em redes sociais sobre um tema específico.
- **Modelos de chatbots para tarefas específicas:** Chatbots treinados para responder a perguntas sobre um determinado produto ou serviço.

## Experimental evidence on the productivity effects of generative artificial intelligence

SHAKKED NOY AND WHITNEY ZHANG [Authors Info & Affiliations](#)

SCIENCE • 13 Jul 2023 • Vol 381, Issue 6654 • pp. 187-192 • DOI: 10.1126/science.adh2586

[science.org/doi/10.1126/science.adh2586](https://science.org/doi/10.1126/science.adh2586)

# GANHOS COM LLMS

## ChatGPT gives an extra productivity boost to weaker writers

The AI program allows people with limited writing skills to create higher-quality texts—but makes little difference to proficient writers' work quality.

[Mariana Lenharo](#)

[nature.com/articles/d41586-023-02270-9](https://nature.com/articles/d41586-023-02270-9)



LIMITAÇÕES

*Alucinações*

- **Fine-tuning:**

- Conceito: É o processo de ajustar os parâmetros de um modelo pré-treinado em um conjunto de dados específico e menor.

- **Few-shot learning:**

- Nesta técnica, o modelo é exposto a poucos exemplos da tarefa desejada antes de ser solicitado a realizar uma nova tarefa. O modelo aprende a generalizar a partir desses poucos exemplos

- **Prompt Engineering:**

- A arte de criar prompts (instruções) precisos e bem elaborados para guiar o modelo na geração de respostas desejadas.

- **Reinforcement Learning from Human Feedback (RLHF)**

- O modelo é treinado a maximizar uma recompensa definida por humanos, geralmente através de interações com humanos. Exemplo: Treinar um chatbot a gerar respostas mais humanas e empáticas através de feedback de usuários.



**ADAPTAR UM  
LLM**



# APLICAÇÕES

## PROCESSAMENTO DE LINGUAGEM NATURAL (PLN)

### CRIAÇÃO DE CONTEÚDO

Os criadores de conteúdo aproveitam grandes modelos de linguagem para gerar artigos, roteiros e materiais de marketing com eficiência

Grandes modelos de linguagem são amplamente utilizados em tarefas de PNL, como geração de texto, análise de sentimentos, tradução de idiomas e chatbots.



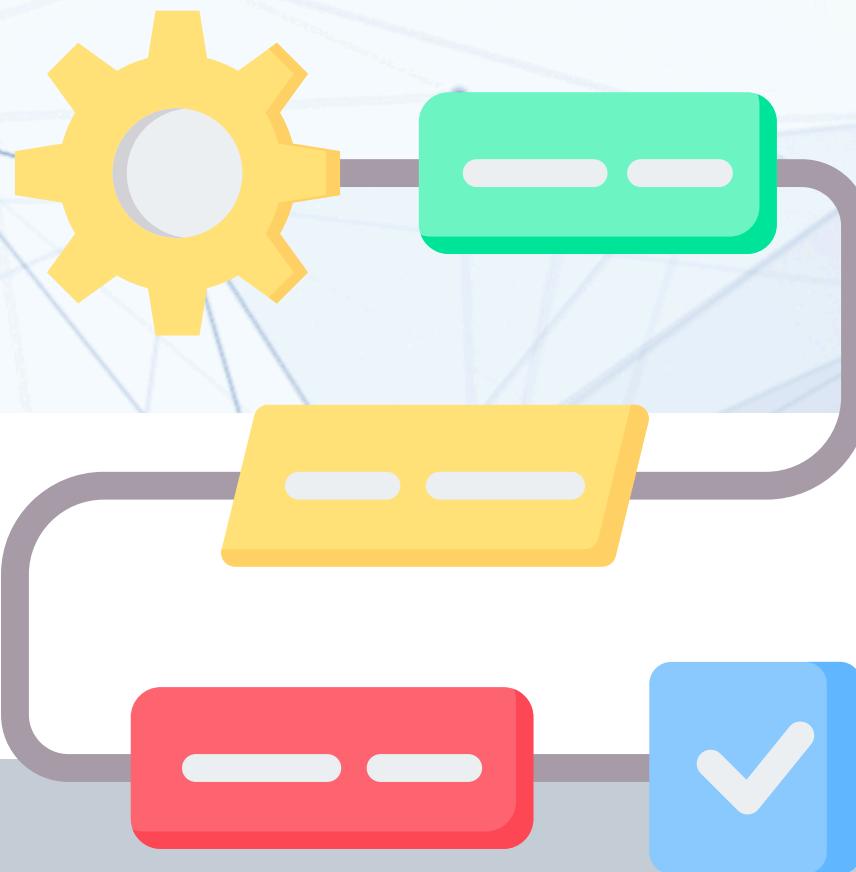
## DIAGNÓSTICO MÉDICO

Modelos de linguagem estão sendo desenvolvidos para auxiliar no diagnóstico médico por meio da análise de sintomas de pacientes e registros médicos

## APLICAÇÕES

### CHATBOTS E ASSISTENTES VIRTUAIS

- Atendimento ao cliente, assistentes pessoais, agendando compromissos, enviando lembretes, controlando dispositivos inteligentes.



## CONSTRUÇÃO DE WORKFLOWS E AUTOMAÇÃO

Realizando tarefas como extração de dados, classificação e organização de informações.

## APLICAÇÕES

### GERAÇÃO DE CÓDIGO

Sugerindo o próximo código a ser escrito, acelerando o desenvolvimento.



# E MUITO MAIS

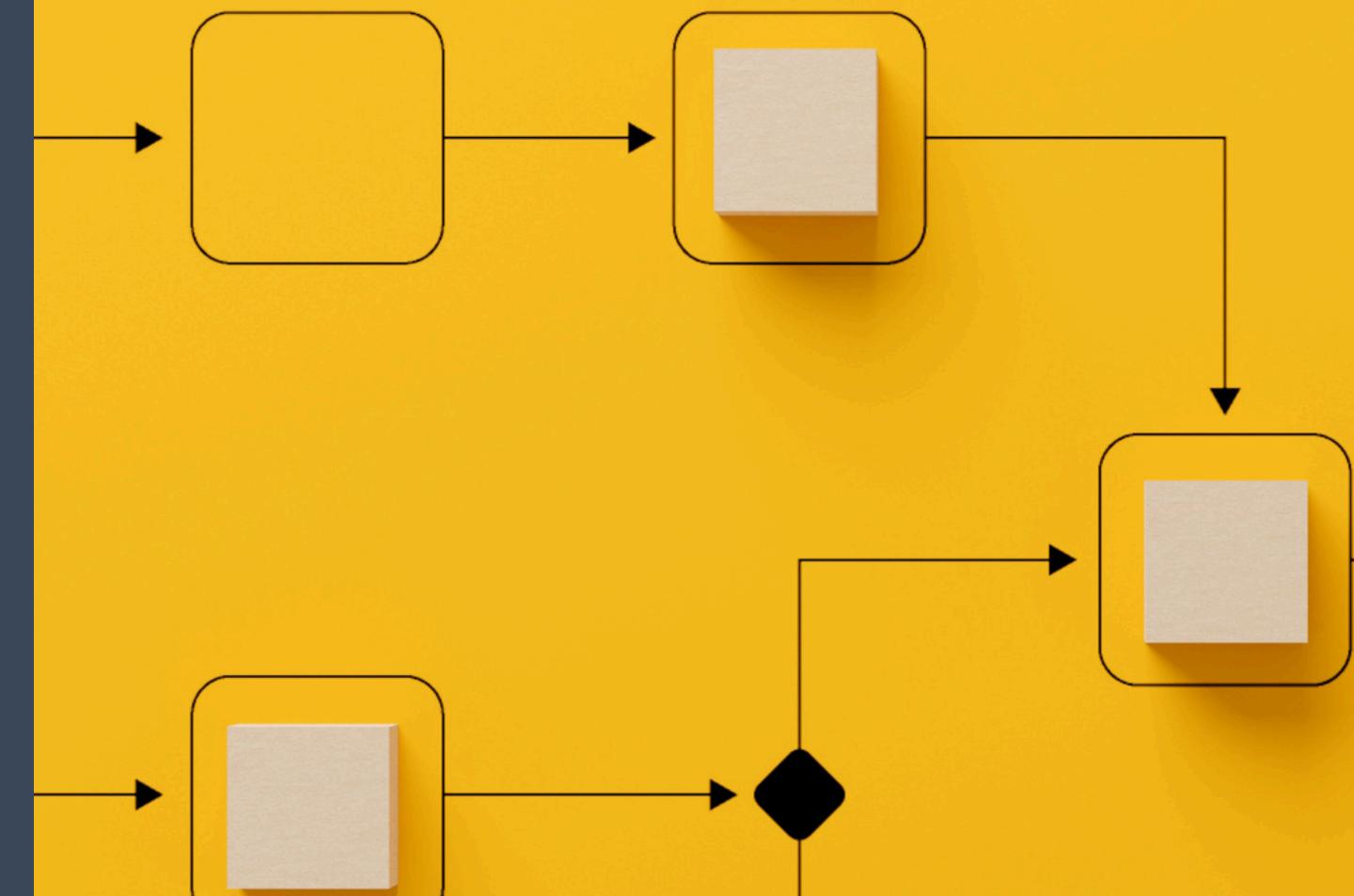
**Educação e Treinamento  
Pesquisa e Desenvolvimento  
Criatividade e Design  
Direito  
Arquitetura  
etc**

- RAG é um acrônimo em inglês que significa **Retrieval-Augmented Generation**. Em português, podemos traduzi-lo como **Geração Aumentada por Recuperação**.

- Essa é uma técnica que combina o poder dos grandes modelos de linguagem (LLMs) com a capacidade de buscar informações em bases de conhecimento externas.

- Como funciona o RAG?

- Consulta: O usuário faz uma pergunta ou solicita uma informação.
  - Recuperação: O sistema busca em uma base de conhecimento (que pode ser um banco de dados, conjunto de documentos ou até mesmo a internet) por informações relevantes à consulta.
  - Geração: O LLM utiliza as informações recuperadas como contexto para gerar uma resposta mais precisa, relevante e informativa.



**RAG**

- Um embedding é um vetor (lista) de números de ponto flutuante. A distância entre dois vetores mede sua relação. Pequenas distâncias sugerem alta relação e grandes distâncias sugerem baixa relação.
- Os embeddings de texto medem a relação de strings de texto. Embeddings são comumente usados para:
  - Pesquisar (onde os resultados são classificados por relevância para uma sequência de consulta)
  - Agrupamento (onde as sequências de texto são agrupadas por similaridade)
  - Recomendações (onde itens com sequências de texto relacionadas são recomendados)
  - Detecção de anomalias (onde são identificados valores discrepantes com pouca relação)
  - Medição de diversidade (onde as distribuições de similaridade são analisadas)
  - Classificação (onde as sequências de texto são classificadas por seu rótulo mais semelhante)

X	1	2	3	4	5	6	7	8
1	1	2	3	4	5	6	7	8
2	2	4	6	8	10	12	14	16
3	3	6	9	12	15	18	21	24
4	4	8	12	16	20	24	28	32
5	5	10	15	20	25	30	35	40
6	6	12	18	24	30	36	42	48
7	7	14	21	28	35	42	49	56

# EMBEDDINGS

ISTO É  
TUDO  
PESSOAL



benevidfelix



benevid@unemat.br