

IA Integrada ao Desenvolvimento e Automação de Software: Da Teoria à Prática

2024
2024
2024

Dia do ACBr



César Cardoso

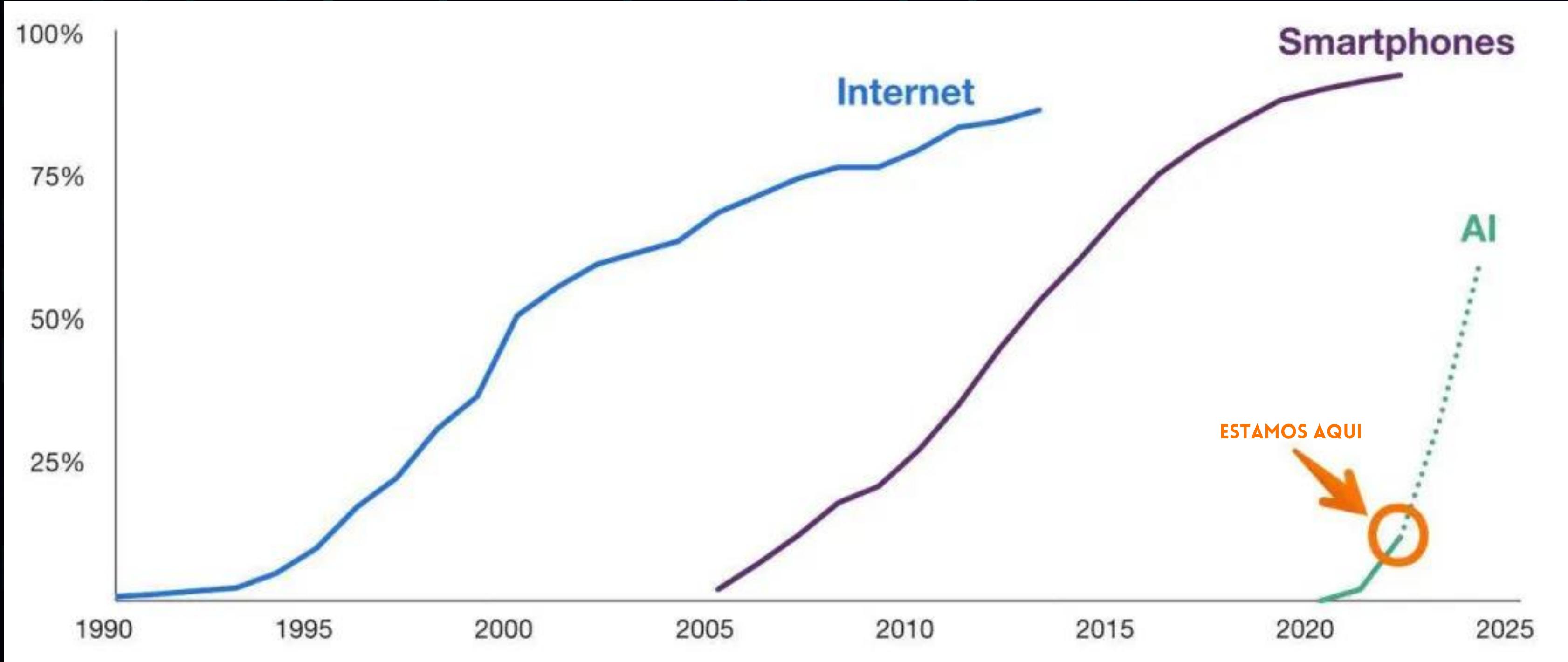
Quem sou eu

- César Augusto Santiago Cardoso
- Graduado em Análise e Desenvolvimento de Sistemas
- Pós-graduado em Desenvolvimento de Sistemas Web
- Desenvolvedor Delphi/PHP desde 2009
- Desenvolvedor Delphi na Solusys Sistemas
- Criador de conteúdo no canal Code4Delphi
- Palestrante



César Cardoso

Uso da IA



Agentes de IA

O futuro da IA - Bill Gates

“Os agentes não apenas vão mudar como todos interagem com computadores. Eles também vão revolucionar a indústria de software, trazendo a maior revolução em computação desde que passamos de digitar comandos para tocar em ícones”

“Você nunca mais vai acessar um site de busca”

“Você nunca mais vai acessar a Amazon”



Linha do tempo da IA

Tecnologia evoluiu até a IA



Clippy - Office 97 até 2001
O assistente da Microsoft
já apareceu na lista das 50
Piores Invenções da Time.

Ele **não** usava IA



akinator®

O jogo faz perguntas e adivinha em quem você está pensando. Essa IA é baseada em um banco de dados com mais de 20 milhões de pessoas, animais e coisas.



Optimus Gen 2 - Humanoide com IA

Linha do tempo da IA

1943

SISTEMA NERVOSO HUMANO

Os cientistas Warren McCulloch e Walter Pitts apresentam um artigo que fala pela 1ª vez de um modelo matemático com base no sistema nervoso humano

1982

REDES NEURAIS ARTIFICIAIS

John Hopfield cria o conceito de redes neurais artificiais. Também chamado de rede Hopfield.

1950

TESTE DE TURING - ALAN TURING

Jogo da imitação. Desenvolvimento de uma forma de avaliar se uma máquina consegue se passar por um humano em uma conversa escrita

1997

DEEP BLUE VENCE NO XADREZ

O campeão soviético de xadrez Garry Kasparov foi derrotado em uma das rodadas pelo computador Deep Blue

1956

TERMO INTELIGÊNCIA ARTIFICIAL

Conferência de Dartmouth, EUA. O cientista da computação John McCarthy cunha o termo Inteligência Artificial

2014

TESTE DE TURING É VENCIDO

Um chatbot chamado Eugene Goostman conseguiu vencer o teste de Turing

1964

ELIZA, O PRIMEIRO CHATBOT DO MUNDO

Conversava de forma automática imitando uma psicanalista. Suas respostas eram baseadas em palavras-chave e estrutura sintática

2017

TRANSFORMERS

Arquitetura de rede neural artificial, lançada num artigo da Google de 2017 intitulado "Atenção é tudo que você precisa". (O T da sigla GPT significa Transformer)

1970

1970 – 1980: INVERNO DA IA

Período sombrio conhecido como inverno da inteligência artificial, com poucas novidades, cortes nos investimentos e baixa atenção ao setor

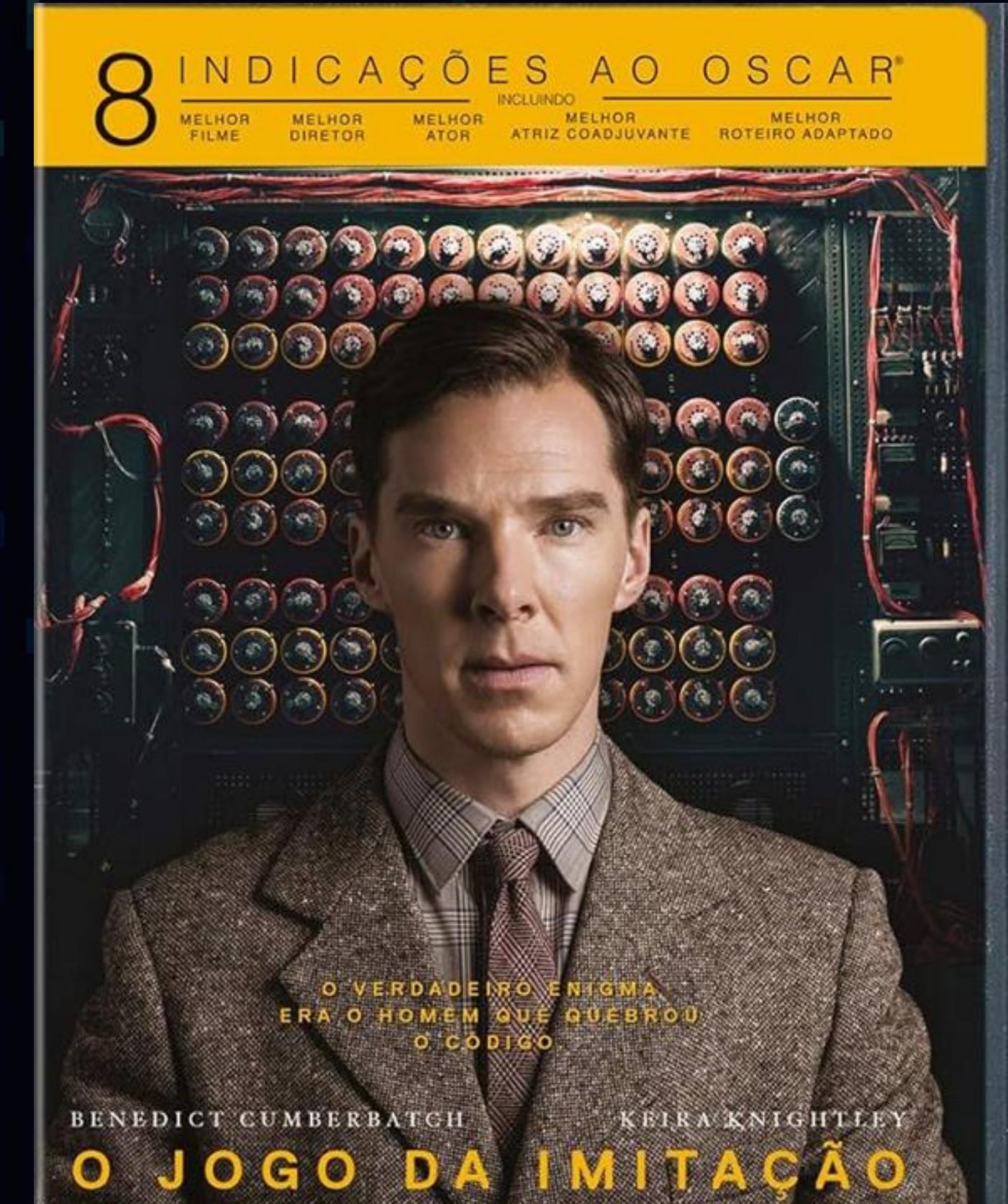
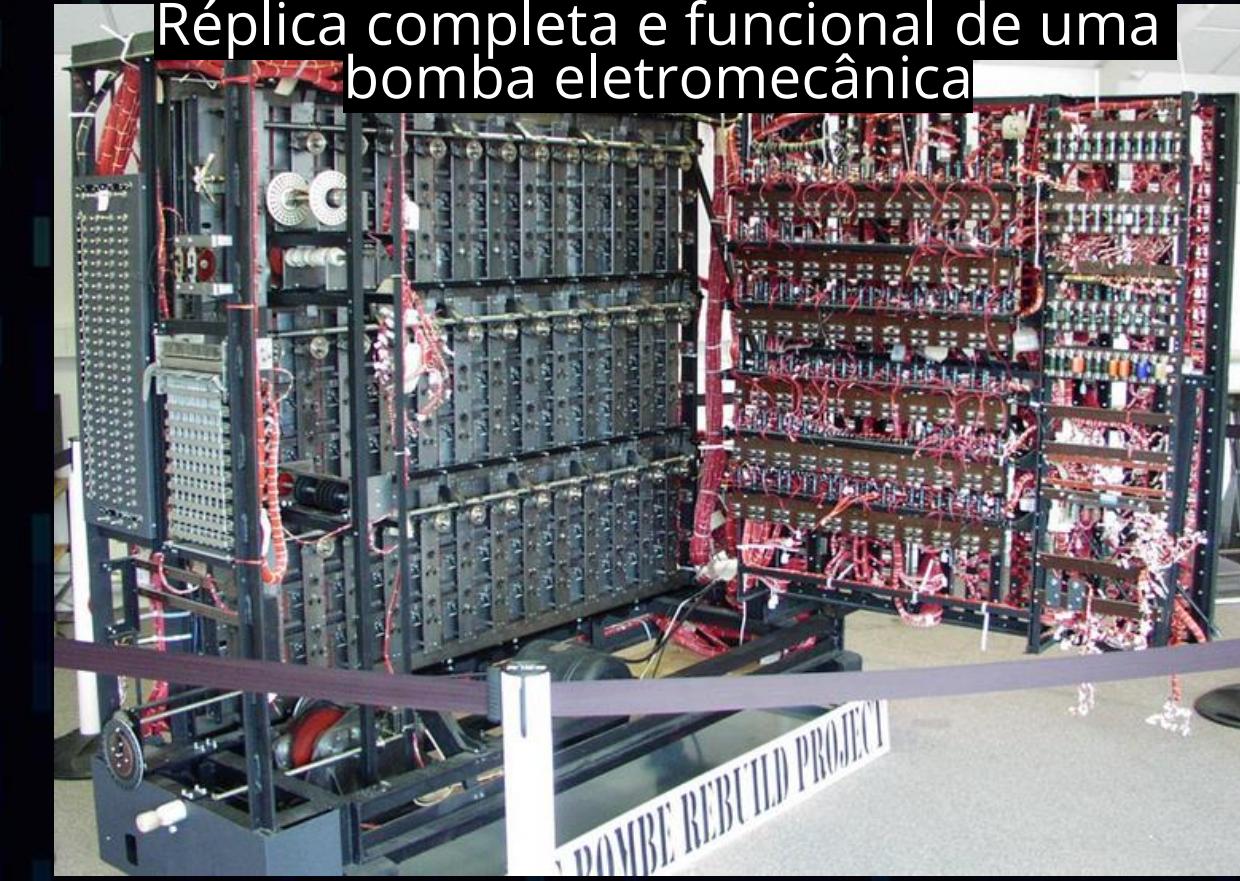
2022

CHAT GPT

Open AI lança Chat GPT, que rapidamente torna-se popular

Linha do tempo da IA

Alan Turing, o pai da computação



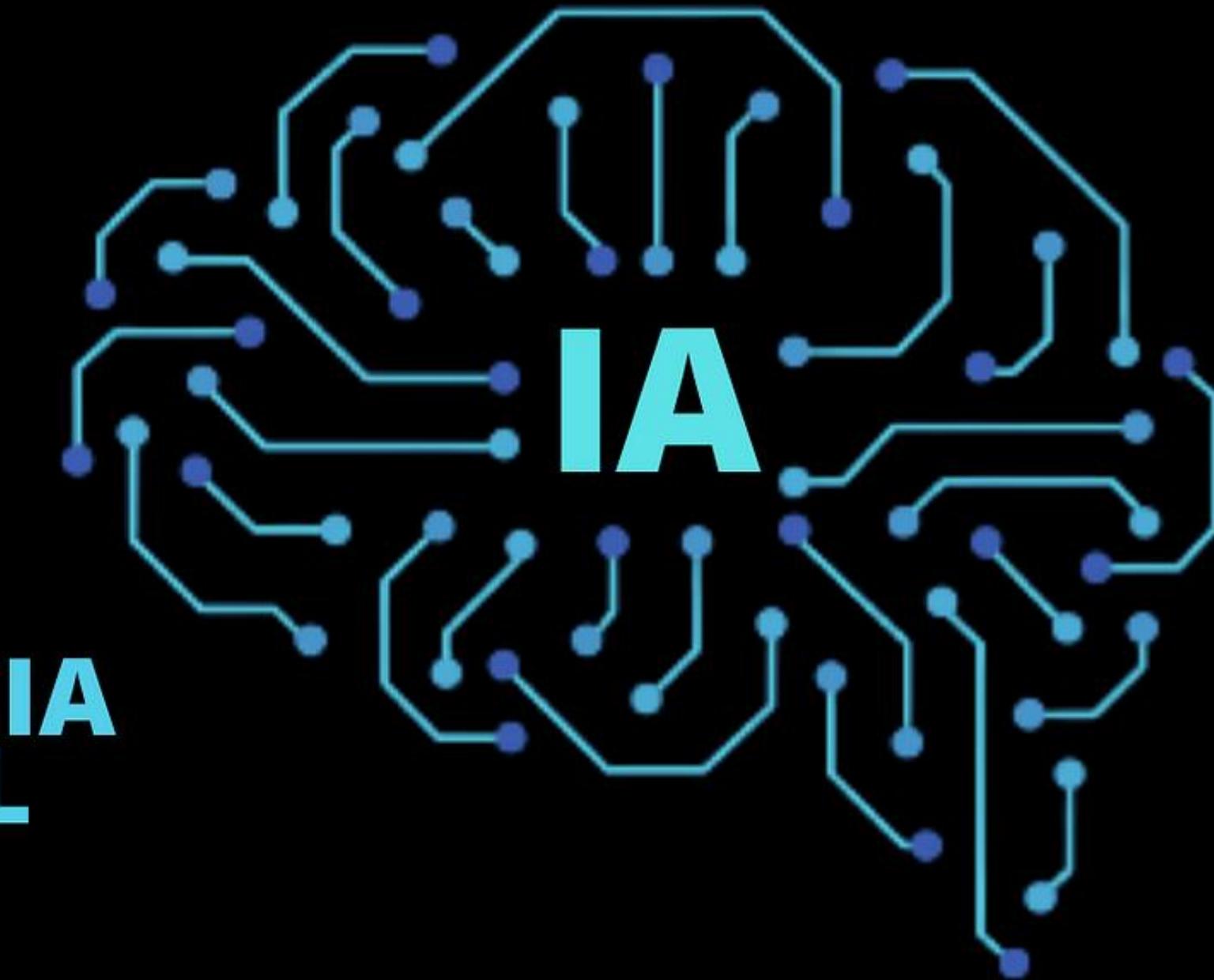
O que é uma IA?

Simulação da inteligência humana feita por máquinas programadas



**INTELIGÊNCIA
ARTIFICIAL**

Você sabe o que é?



IA Generativa

Têm a capacidade de criar novos conteúdos a partir de conjuntos de dados pré-existentes, como texto, imagens, música, áudio e vídeos.

O que a IA generativa pode criar

- Texto
- Imagens
- Sons
- Músicas
- Áudios
- Vídeos
- Códigos

Modelos de IA Generativa

Exemplos de IA generativa

IA para
Textos / Chat



GPT **OpenAI**

Gemini **Google**



Open Source

Llama **Facebook**

Grok **Twitter (X)**

IA para
Imagens



DALL-E 3

S.

Stable Diffusion



Midjourney

IA para
Áudio



TTS

OpenAI
Whisper

Eleven
Labs

IA para
Vídeos / Avatares



R **runway**



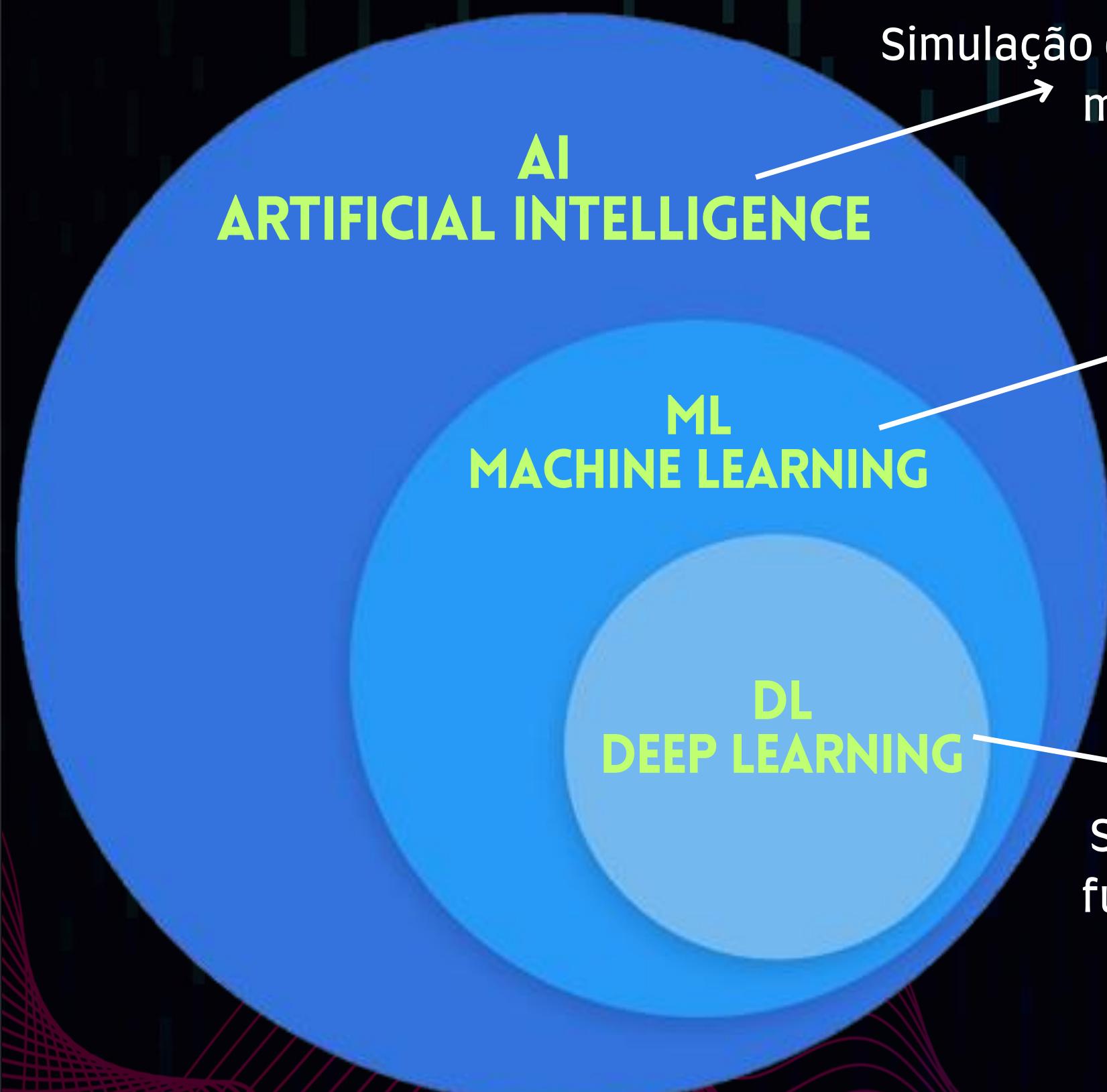
PIKA LABS

S.

Stable Diffusion

Machine Learning e Deep Learning

Conceitos



Simulação da inteligência humana feita por máquinas programadas

Aprendizado de máquina

Subcampo da IA no qual as máquinas aprendem com conjuntos de dados em vez de serem explicitamente programadas.

Termo de Arthur Samuel em 1959

Principais tipos do ML

- Aprendizado supervisionado
- Aprendizado não supervisionado
- Aprendizado por reforço

Aprendizado profundo

Subcampo da ML que realiza a simulação do funcionamento do cérebro humano para que as máquinas possam aprender e executar tarefas complexas

O que são LLMs?

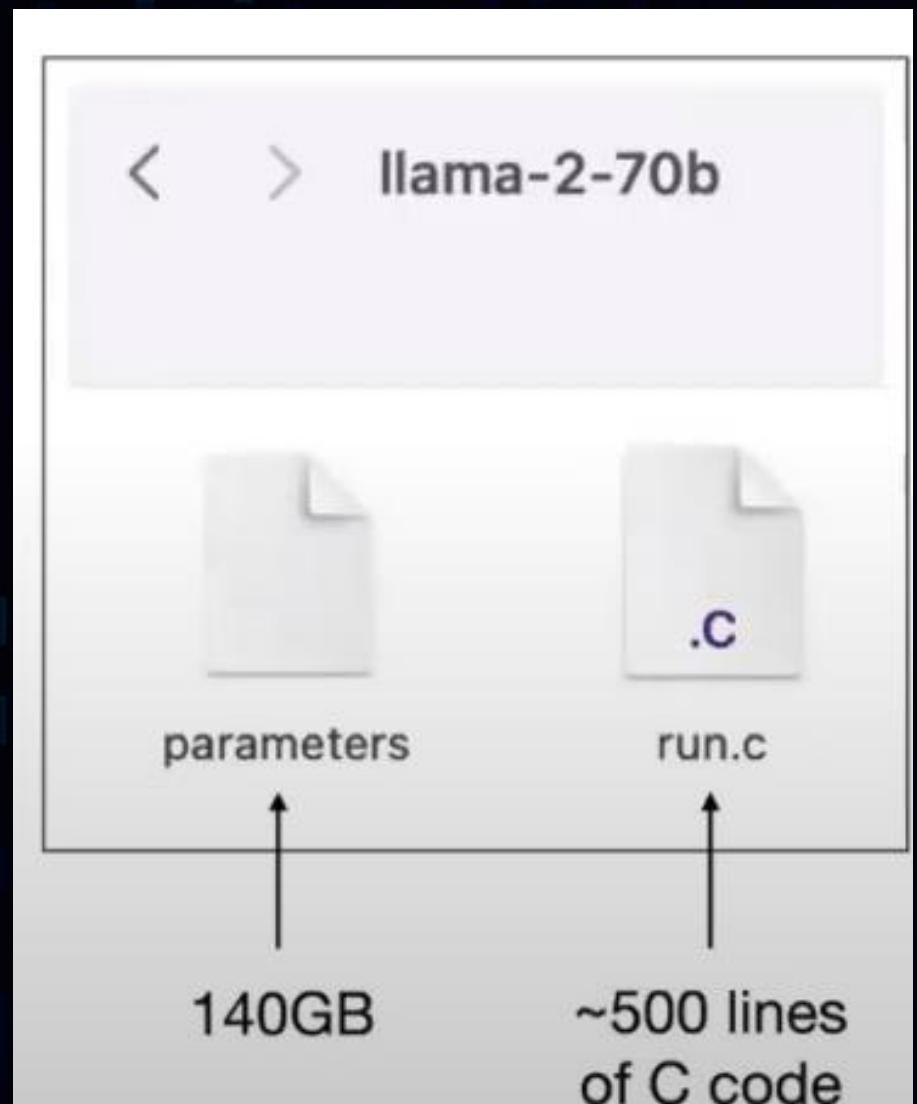
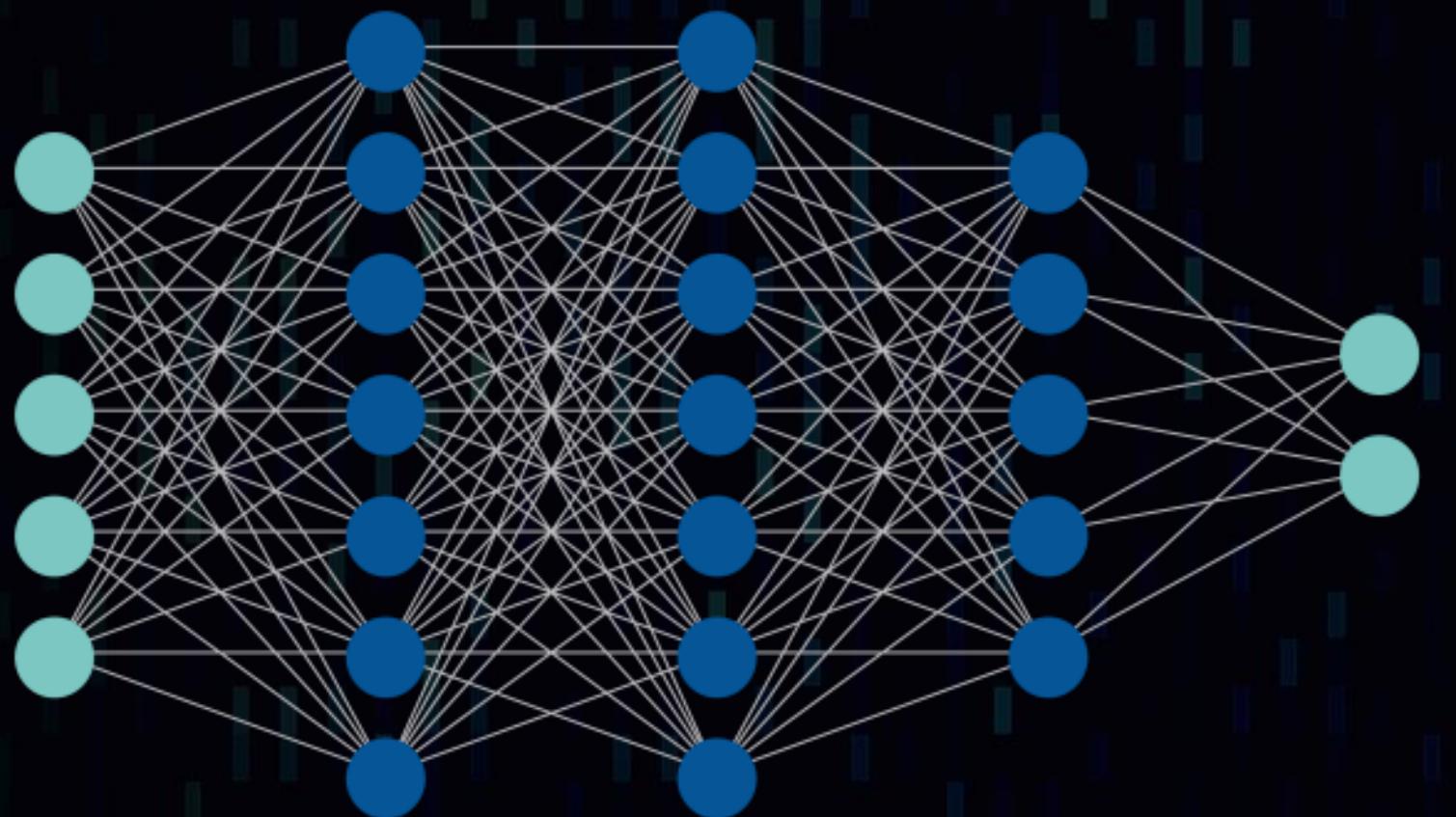
Grandes modelos de linguagem / Modelos de Linguagem de Larga Escala



O que são LMs?

Podemos resumir LLM como sendo 2 arquivos em um diretório

Parâmetros



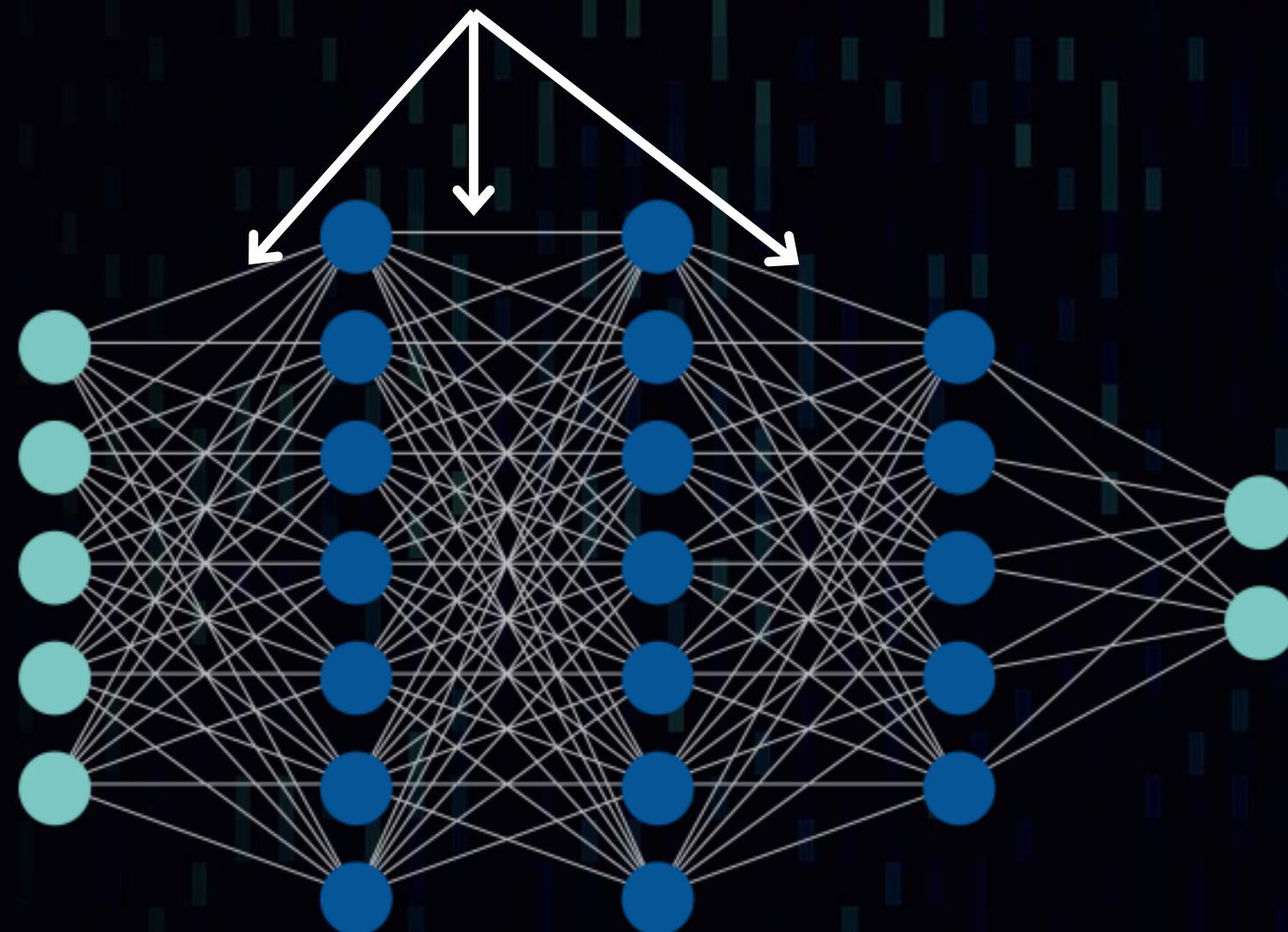
Códigos



O que são parâmetros?

Relação entre palavras

Parâmetros



É a relação entre as palavras (token) e qual é a relação e influência de uma palavra sobre a outra em um determinado contexto (tudo vira token vira números)

Tokenização de textos

Os LLMs não processam textos e sim tokens

Tokenização é o processo de dividir um texto em unidades menores chamadas tokens.

Ferramenta: <https://platform.openai.com/tokenizer>

The image shows two side-by-side screenshots of the OpenAI Tokenizer interface. Both screenshots have a dark background with purple wavy lines on the sides. At the top, there are two buttons: 'GPT-3.5 & GPT-4' (highlighted in green) and 'GPT-3 (Legacy)'. Below the buttons is a text input field containing the text 'O ACBr é'. Underneath the input field are two buttons: 'Clear' and 'Show example'. Below these buttons is a table with two columns: 'Tokens' and 'Characters'. The 'Tokens' column shows the number '4' in blue, and the 'Characters' column shows the number '8' in red. At the bottom of the interface, there are two buttons: 'Text' (highlighted in green) and 'Token IDs'. In the 'Text' view, the input 'O ACBr é' is shown with each character in a separate colored box (purple for 'O', green for 'A', yellow for 'C', blue for 'B', red for 'r', and purple for 'é'). In the 'Token IDs' view, the input is shown as a list of four numbers: [46, 10807, 6971, 4046].

De onde vem os parâmetros?

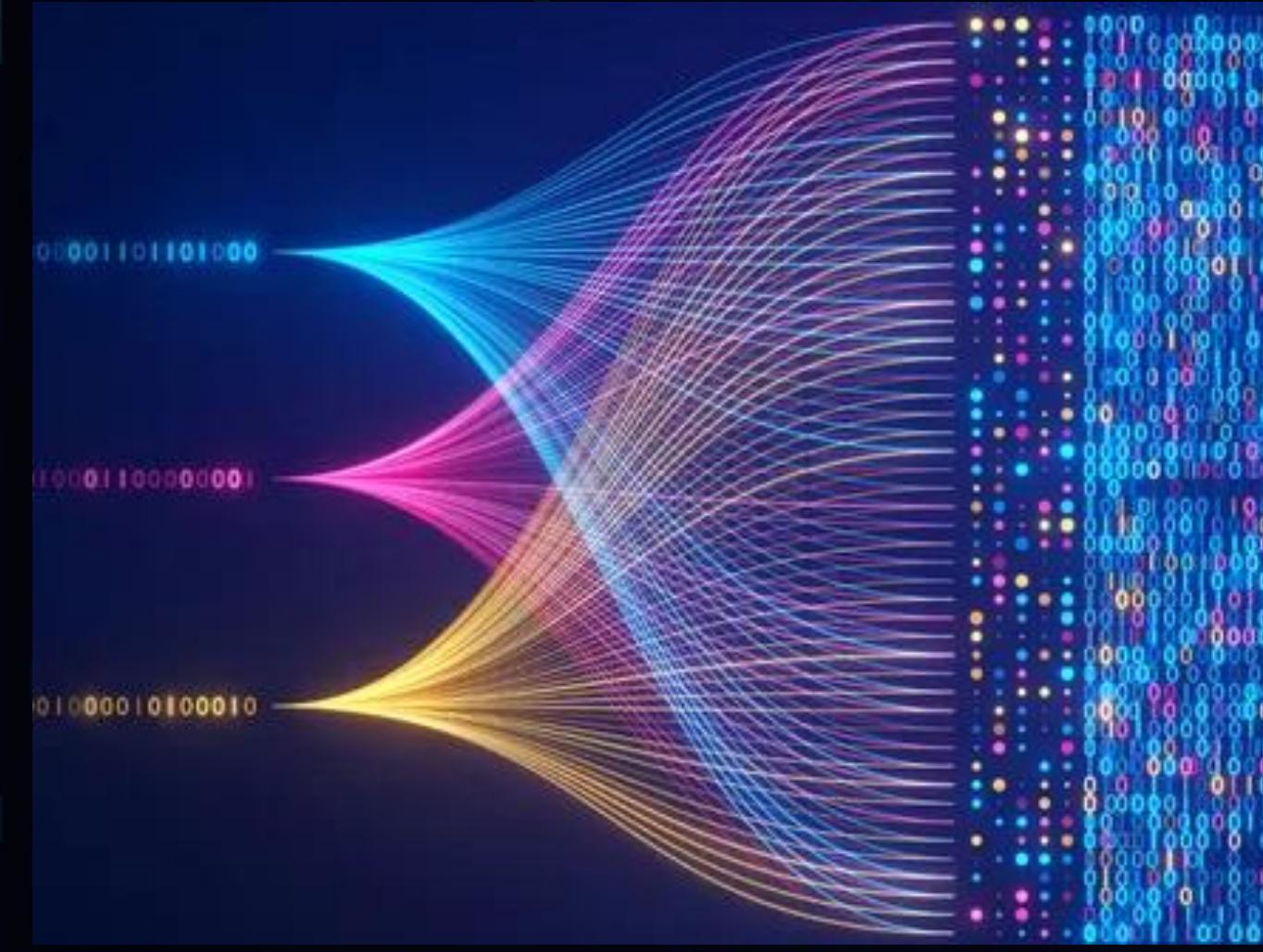
Dados do LLama 2



Textos (internet, livros, etc.)
~10 TB de textos



6000 GPUs rodando por 12 dias
~\$2M



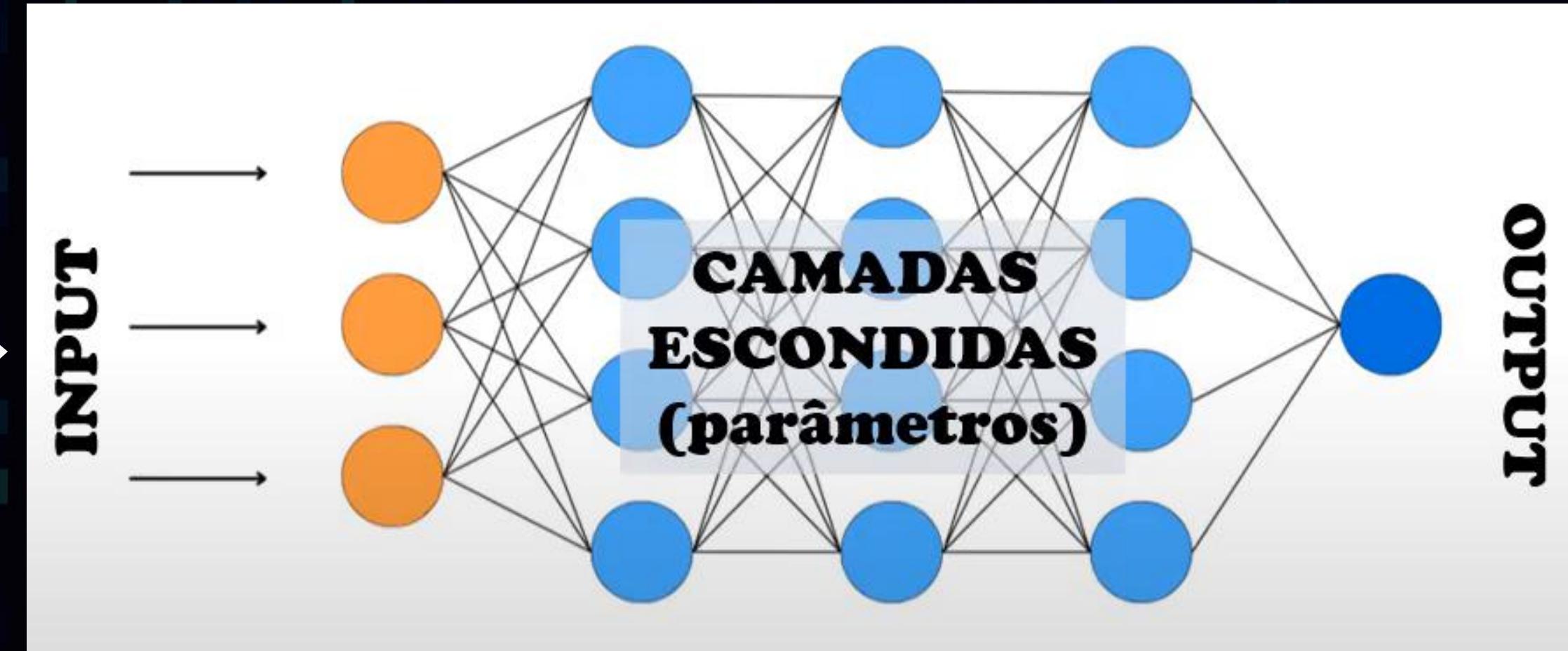
Parâmetros ~140GB
(compressão com perda)

Llama 3: conjunto de dados de treinamento é sete vezes maior que o usado para o Llama 2 e inclui quatro vezes mais código

Rede neural IA

Função primária: prever a próxima palavra da sequência

O gelo é? →



“Alucinação” das IAs

Alucinações das IAs ocorrem com uma frequência que varia de 3% e 27%

ChatGPT 3.5 ▾



You

Qual a maior palavra de 5 letras do vocabulário português?



ChatGPT

A maior palavra de 5 letras no vocabulário português é "esque", que significa "lembrança" ou "memória".



You

Tente de novo



ChatGPT

Peço desculpas pelo erro anterior. A maior palavra de 5 letras no vocabulário português é "esquer".



You

Mas Esquer não tem 5 letras



ChatGPT

Você está absolutamente correto, e peço desculpas pelo erro novamente. A maior palavra de 5 letras no vocabulário português é "quest".



You

Quest é uma palavra do vocabulário português?



ChatGPT

Você está correto, "quest" não é uma palavra do vocabulário português. Peço desculpas pelo engano. A palavra correta é "queij".



ChatGPT preso?

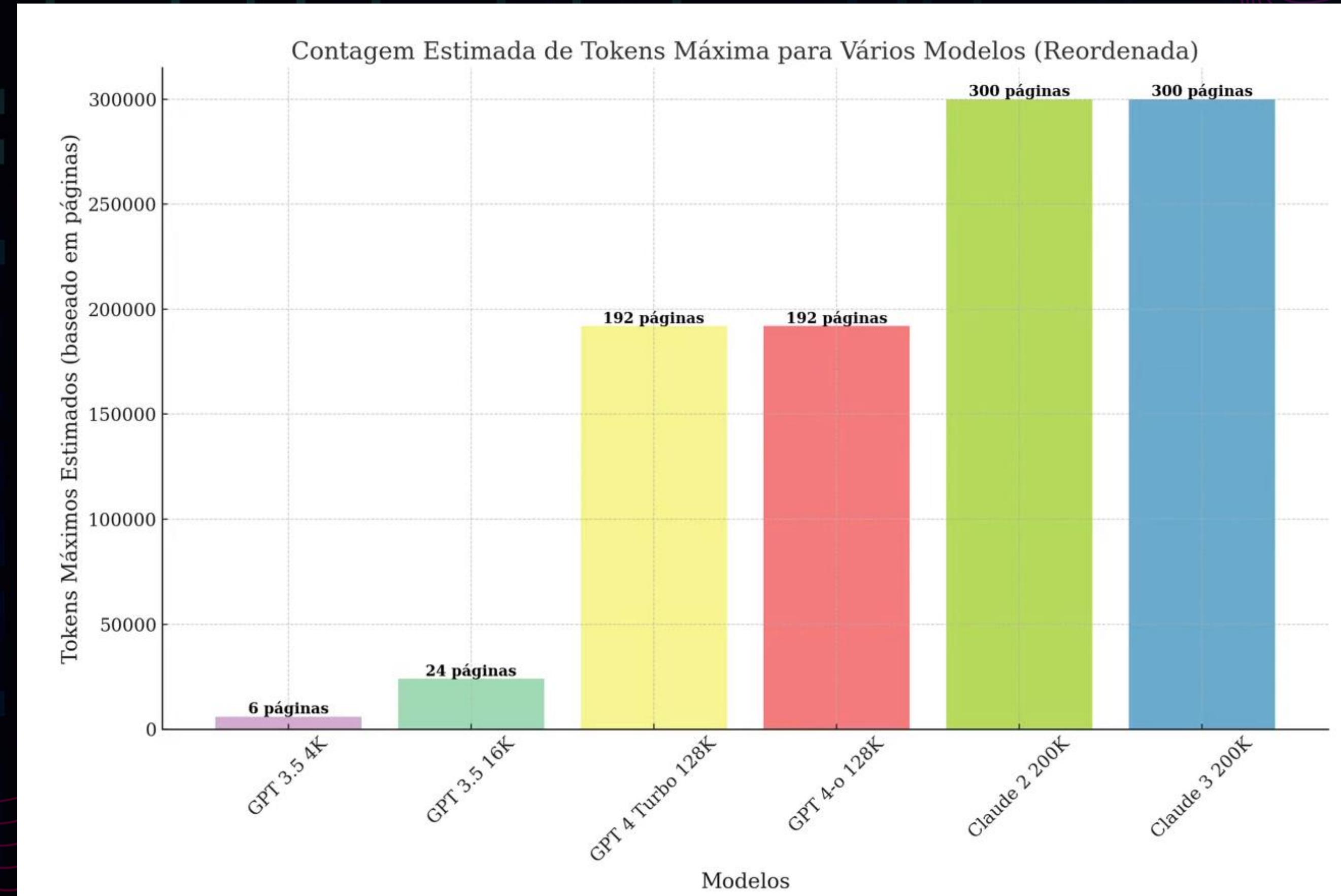
Curiosidades

- ChatGPT inventa lavagem de dinheiro e é **alvo de 1ª ação na Justiça** nos EUA ([Link](#))
- Juiz federal usa ChatGPT e IA inventa informações ([Link](#))
- ChatGPT inventa casos para auxiliar advogado em defesa e é descoberto nos EUA ([Link](#))
- ChatGPT inventa acusação de assédio sexual contra renomado professor no EUA ([Link](#))
- Já tive experiência com isso, ao pedir ajuda para ia sobre OTA, **ela inventa interfaces** que não existem

Janela de contexto

ChatGPT

- Janelas de contexto é a capacidade do modelo de prestar atenção no texto adicionado.



Promp e Engenharia de Promp

O que é um prompt?

- Um prompt é um texto em **linguagem natural** que solicita que a IA generativa execute uma tarefa específica. É a pergunta de enviamos.

O que é engenharia de prompt?

- É a prática de projetar melhores perguntas (prompts) para obter uma melhor resposta dos modelo de IA

Engenharia de Promp

Técnicas

Zero-shot

- Envolve **fornecer um comando simples** para gerar uma resposta útil sem que seja necessário adicionar exemplos ou um contexto específico
 - * **Exemplo:** “Descreva como usar um ERP”

Few-shot

- **Fornecer exemplos** para ajudar o modelo a formular uma resposta mais próxima do resultado esperado. Como quais módulos o ERP possui, quais documentos fiscais ele emite, entre outros.
 - * **Exemplo:** “O ERP que possui um módulo fiscal com NF-e, NFC-e e CT-e”
“Com base nesta classe de clientes crie uma classe de produtos”

Engenharia de Promp

Técnicas

Chain-of-Thought (cadeia de pensamento)

- Técnica que treina o modelo para dividir a resposta em diversas etapas menores, como se fosse um passo a passo do raciocínio usado até chegar à resposta.

* **Exemplo:** Para a documentação de emissão de NF-e: Comece sugerindo o cadastro das configurações do certificado. Depois, as configurações do web service. Em seguida, a preencher os dados do emitente da NF-e...

Self-consistency (Autoconsistência)

- Pedir a um modelo o mesmo prompt várias vezes e usa o resultado da maioria como a resposta final. É uma continuação da Chain-of-Thought (cadeia de pensamento) e é mais poderosa quando usada em conjunto com ela.

Engenharia de Promp

Técnicas

Tree of Thought (árvore do pensamento)

- Aprimora o raciocínio da IA ao estruturar respostas em forma de árvore, explorando múltiplos caminhos antes de responder.

* **Exemplo:** Você é um assistente de emissão de Notas Fiscais Eletrônicas (NF-e). Para emitir uma NF-e corretamente, siga estes passos:

1. Verifique o cadastro do cliente: O cliente está cadastrado corretamente no sistema?
2. Calcule os impostos: Os impostos estão sendo calculados corretamente com base nos produtos e na legislação vigente?
3. Confirme os dados dos produtos: Todos os produtos estão descritos corretamente, com códigos e preços adequados?
4. Revise os dados fiscais: Todos os dados fiscais estão completos e corretos, incluindo informações de transporte e pagamento?

Explore cada um desses passos e indique se todos estão corretos ou se há ajustes a serem feitos.

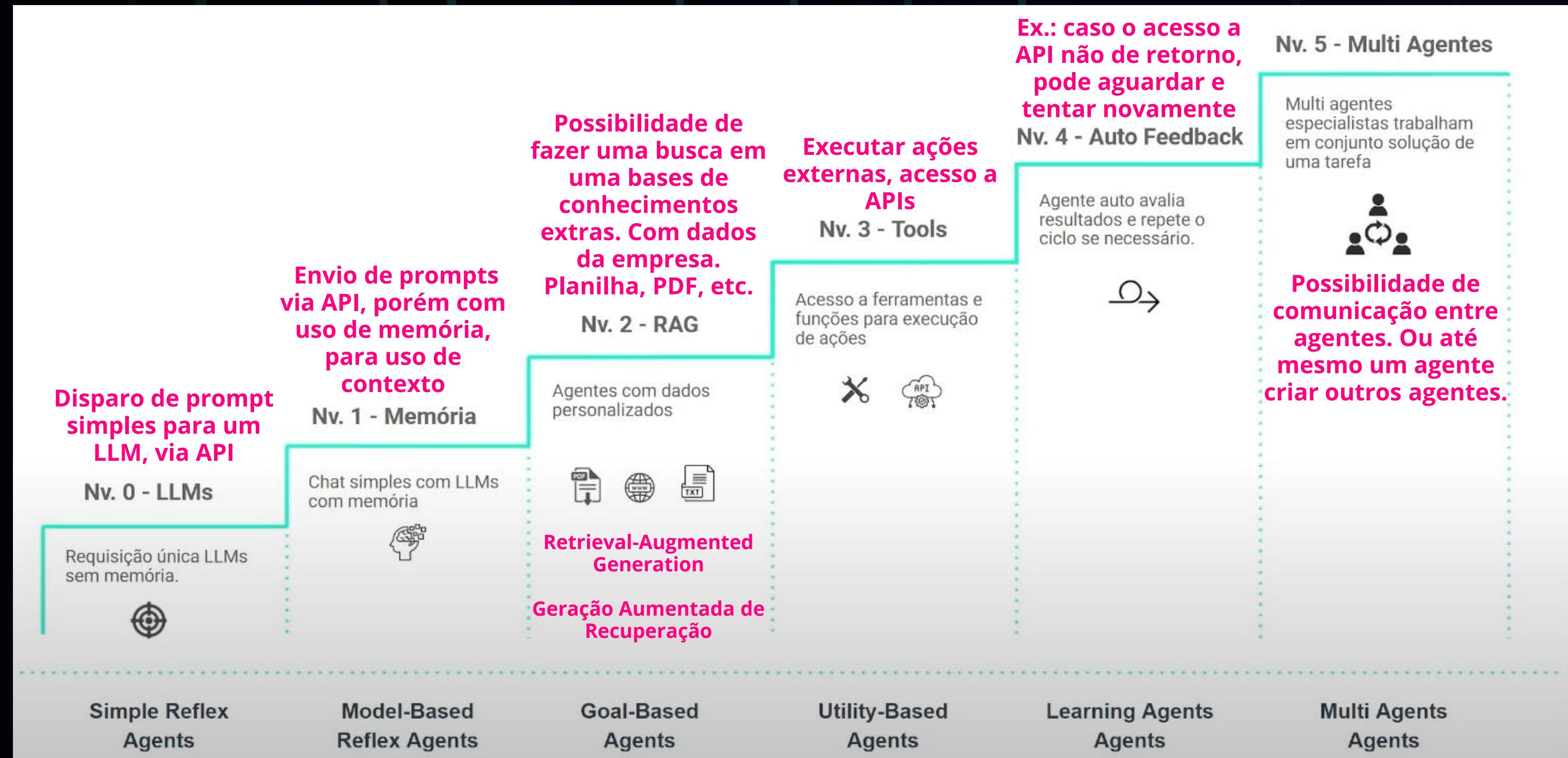
Agentes de IA

O futuro da IA

- Um agente é um **sistema** que utiliza IA para executar tarefas específicas, responder perguntas e **automatizar processos**

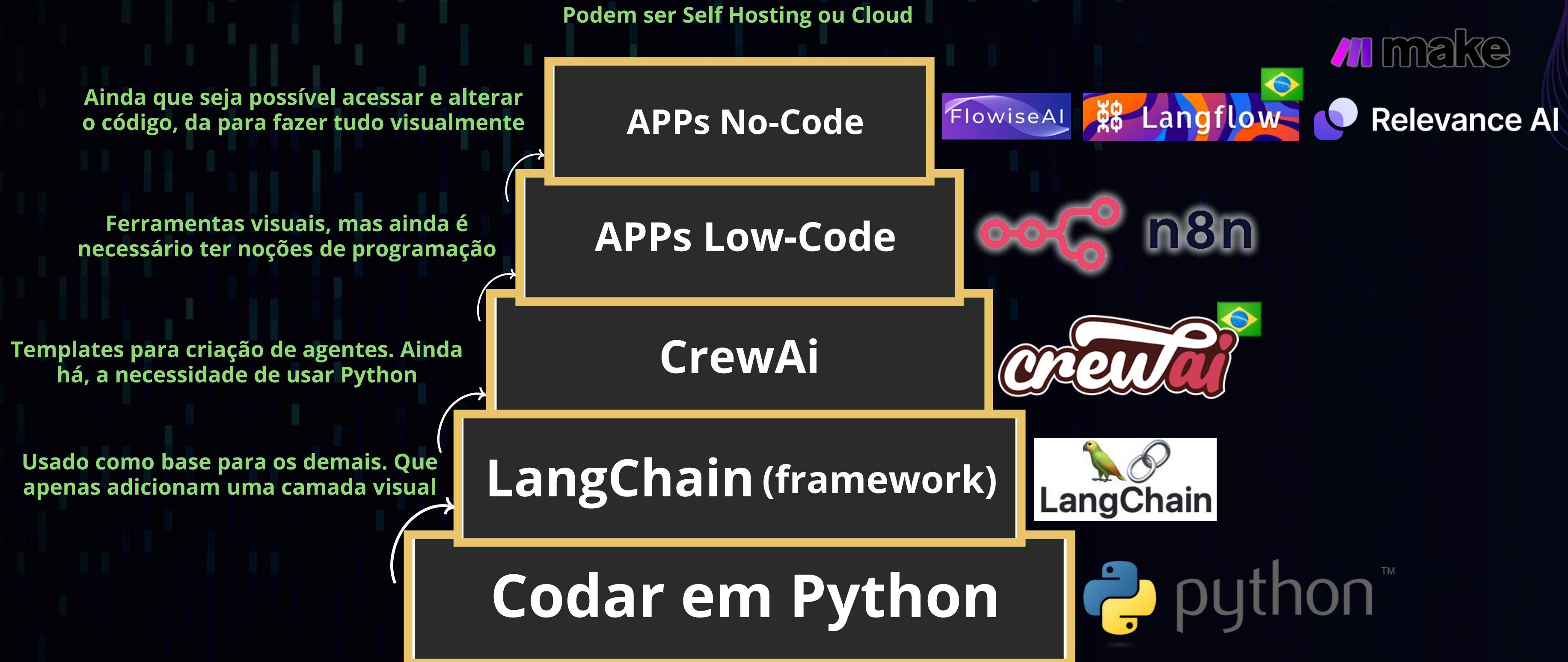
Agentes de IA

Tipos / níveis de Agentes



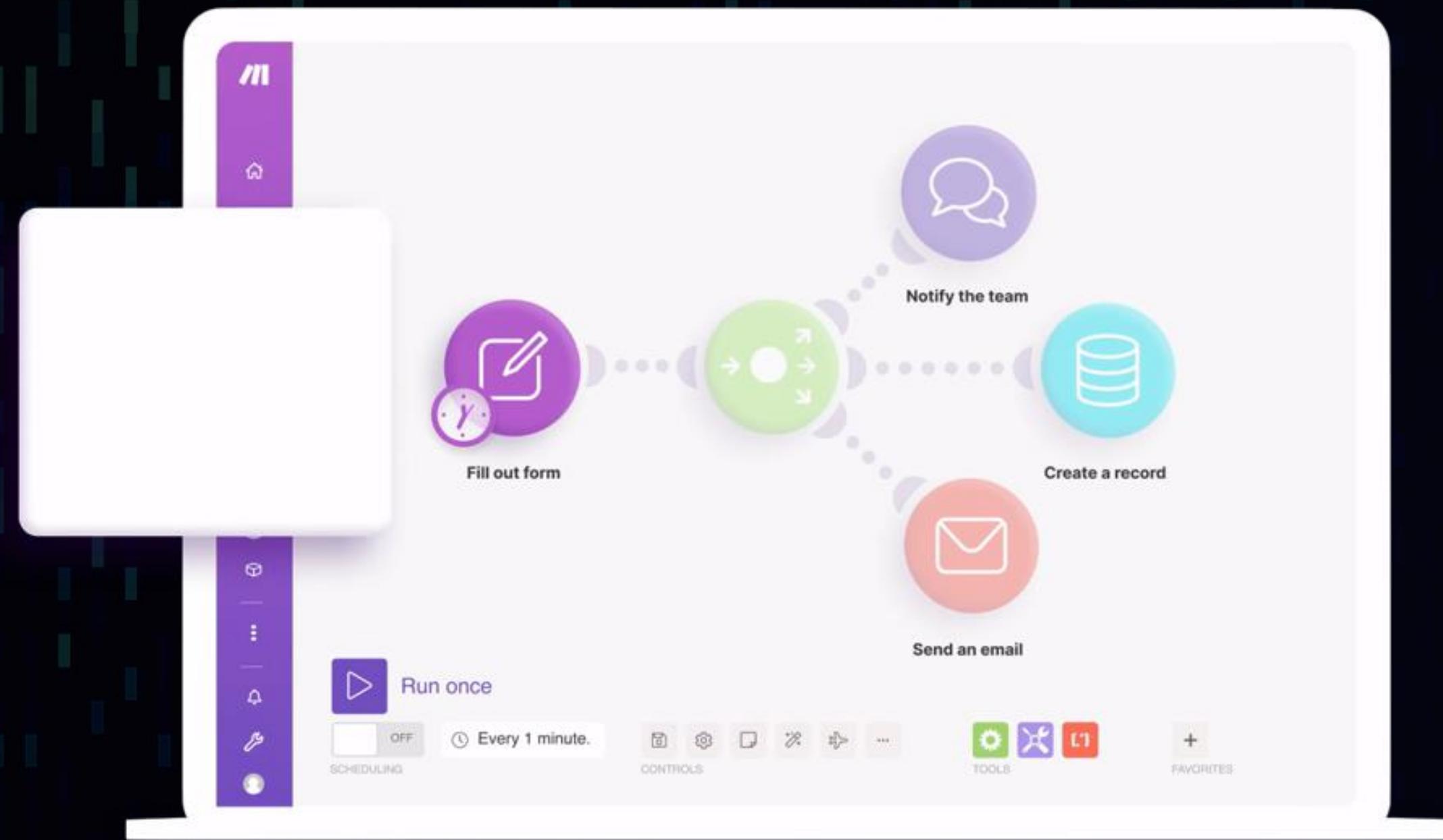
Agentes de IA

Como são desenvolvidos os agentes – Degraus de dificuldade



Agentes de IA

Make



Agentes de IA

Langflow

The screenshot displays the Langflow interface, a low-code platform for building AI applications. The top navigation bar includes 'Demo' (dropdown), 'My Collection' (home icon), 'Community Examples' (magnifying glass icon), 'Star' (star icon) with the value '12889', social sharing icons for Twitter and GitHub, and a settings icon.

The left sidebar lists various components categorized under 'Agents', 'Chains', 'Custom', 'Loaders', 'Embeddings', 'LLMs', 'Memories', 'Output Parsers', 'Prompts', 'Retrievers', 'Text Splitters', 'Toolkits', 'Tools', 'Utilities', 'Vector Stores', and 'Wrappers'. A search bar is also present in the sidebar.

The main workspace shows a flowchart with several nodes:

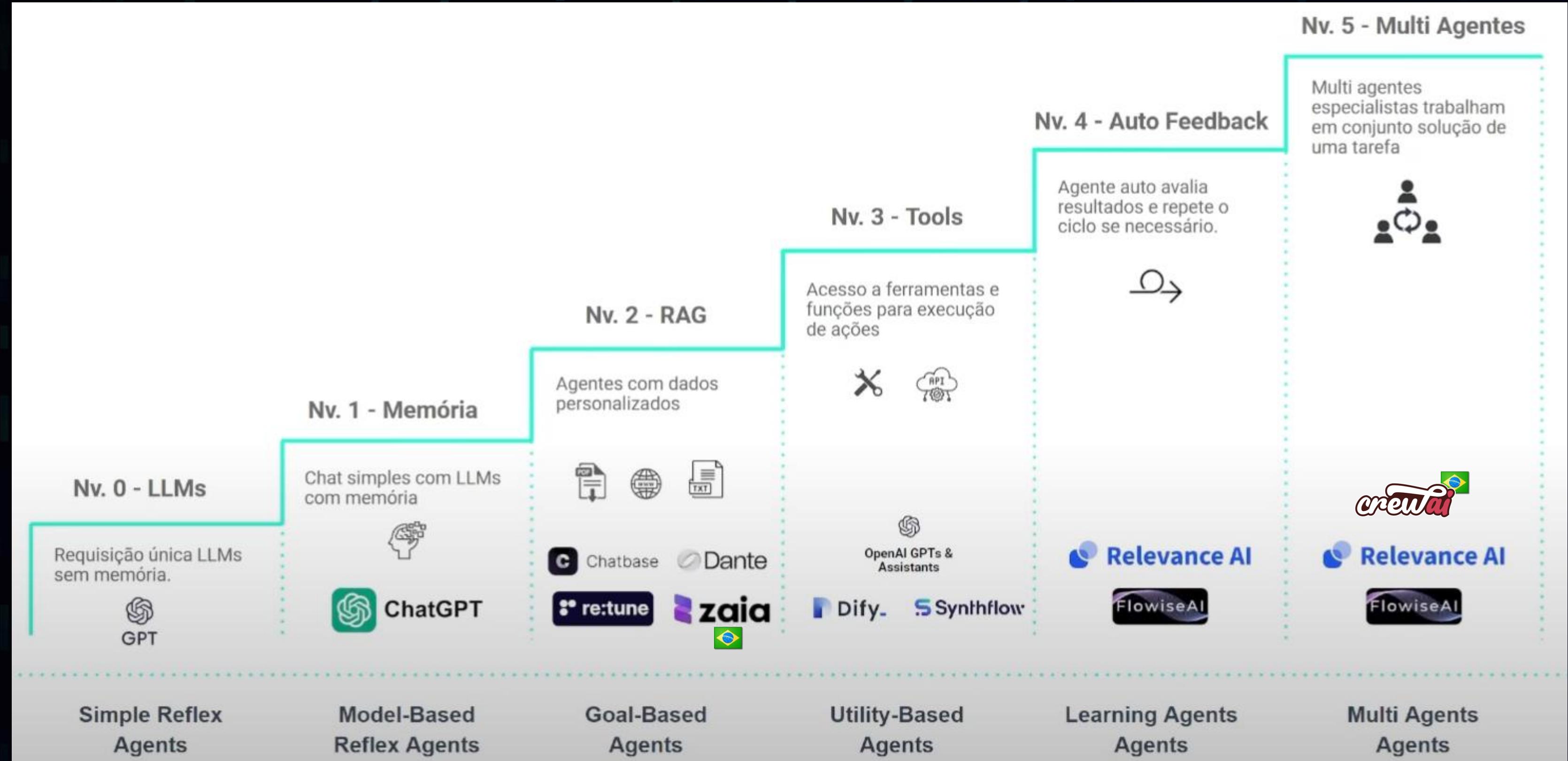
- A 'Feature Description' node (text input component) connects to a 'CustomComponent' node.
- The 'CustomComponent' node connects to an 'LLM Runner' node (beta).
- The 'LLM Runner' node connects to a 'PromptTemplate' node (beta).
- The 'PromptTemplate' node has an input field labeled 'desc' with the placeholder 'Type something'.

A central modal window titled 'Chat' provides instructions for setting up input variables and interacting with agents and chains. It includes sections for 'Input Variables' (with a 'context' variable defined) and a 'Display Prompt' section where a user asks about adjusting investment strategy in light of a global economic crisis. The AI response discusses diversifying the portfolio, focusing on defensive sectors, and investing in companies with strong balance sheets.

On the right side of the workspace, there's a preview area for an 'LLMChain' component, which is described as a chain to run queries against LLMs. It includes fields for 'LLM', 'Memory', and 'Prompt'.

Agentes de IA

Principais ferramentas para criação de agentes



ChatGPT

Principais versões

GPT-1: 117 milhões de parâmetros

GPT-2: 1,5 bilhão de parâmetros

GPT-3

- Lançado em 11 de junho de 2020
- 175 bilhões de parâmetros
- Unimodal (somente texto)
- Tem informações até janeiro de 2022

GPT-4

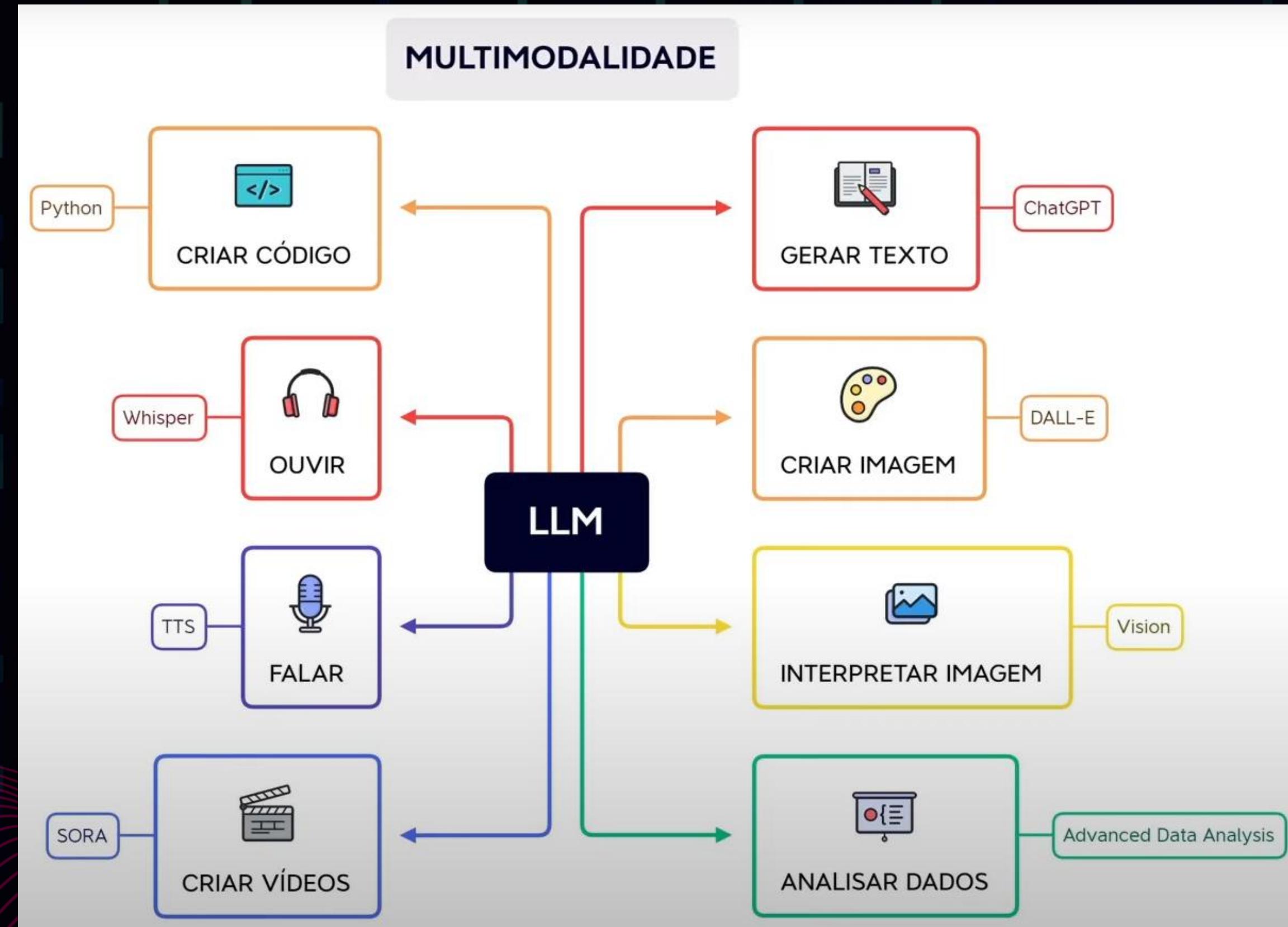
- Lançado em 14 de março de 2023
- Mais de 1 trilhão de parâmetros
- Multimodal (texto e imagens)
- GPT-4 Turbo, dados atualizados até abril/2023
- Custa USD \$20/month

GPT-4o (omni)

- Lançado em 13 de maio de 2024
- Multimodal (texto, criar e interpretar imagens, falar, ouvir, analisar dados, criar vídeos)

ChatGPT

Nas novas versão (4 e 4o pagas) ele se tornou um ecossistema



IA On-Line

Nome	Criador/mantenedor	Lançamento	Link
ChatGPT	OpenAI/Microsoft	Nov / 2022	https://chat.openai.com/
Gemini	Google	Dez / 2023	https://gemini.google.com/
Bard	Google	Julho / 2023	
Gemma	Google	Fev / 2024	https://cloud.google.com/use-gemma?hl=pt-br
Mistral	Mistral AI	Maio / 2023	https://mistral.ai/
Microsoft Copilot	Microsoft	Nov / 2023	https://www.microsoft.com/pt-br/microsoft-copilot
LLama	Meta	Abril / 2023	https://llama.meta.com/llama3/
Claude	Anthropic	2022	https://www.anthropic.com/clause
Amazon	Amazon		Alexa, Bedrock, Titan Text, Anthropic Claude 3
Apple	Apple	Junho / 2024	Apple Intelligence

IA Off-Line Localmente

Ollama é uma ferramenta de código aberto que permite executar e gerenciar LLMs diretamente na sua máquina local. (Docker das IAs)

1º Realize o [Download](#) e a instalação do Ollama

2º Acesse via API: <http://127.0.0.1:11434>



Ollama

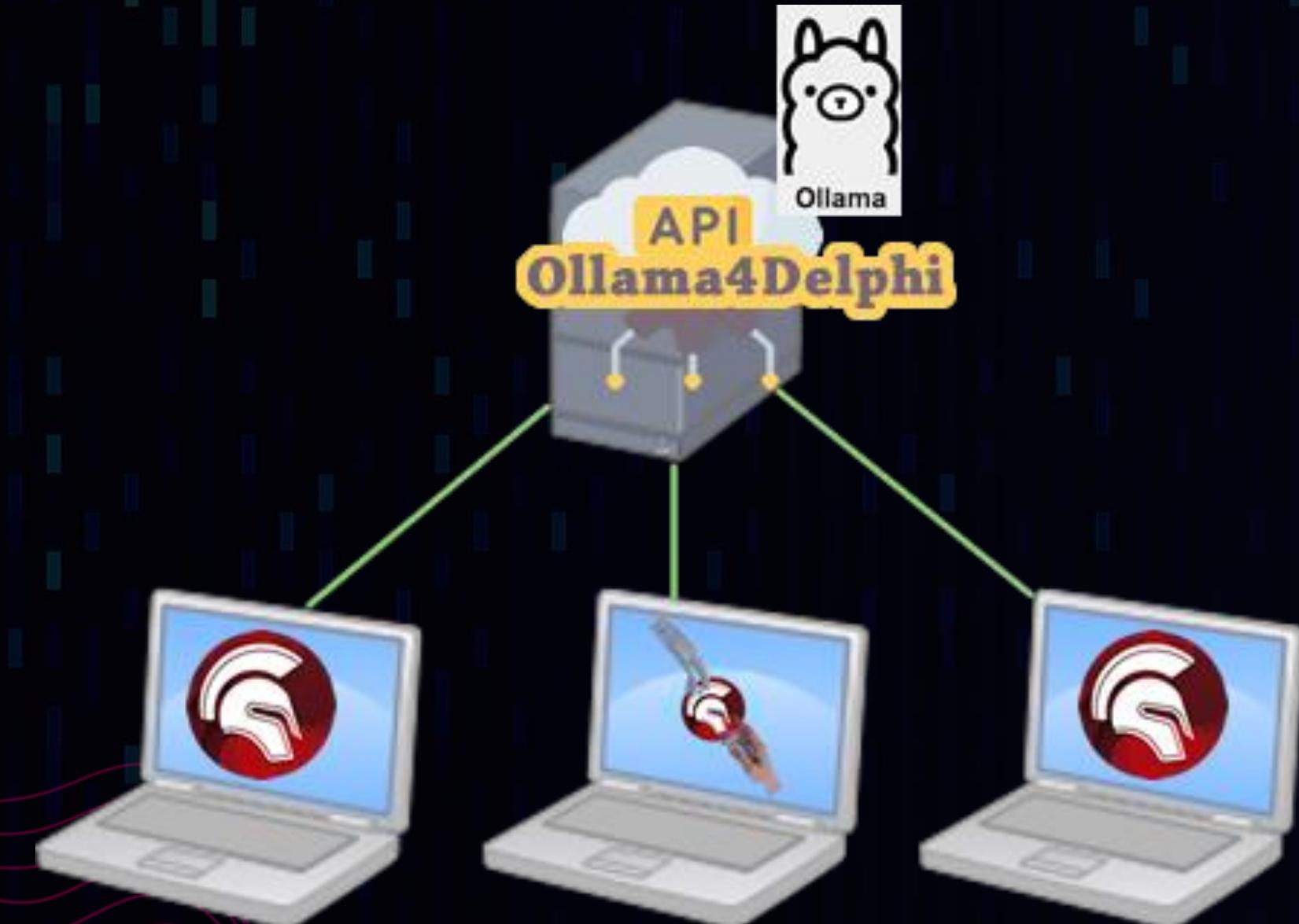
The screenshot shows the Ollama website. On the left, there's a 'Download Ollama' section with links for macOS, Linux, and Windows. A red arrow points from the text '1º acessar página de Models' to the 'Models' tab in the navigation bar. On the right, under the 'Models' section, there's a list of available models. A red arrow points from the text '2º Selecionar um modelo' to the 'llama3.1' model entry. The 'llama3.1' entry includes a description, download statistics (3.3M Pulls, 95 Tags), and a command-line interface (CLI) section with the command 'ollama run llama3.1' highlighted with a red box.

The screenshot shows the GitHub repository for the 'llama3.1' model. It displays basic repository information like 3.3M Pulls, 95 Tags, and an updated date of 3 days ago. Below this, it shows detailed model metadata: model (llama), arch (llama), parameters (8.03B), quantization (Q4_0), size (4.7GB). It also lists params, license, and template. A red box highlights the CLI command 'ollama run llama3.1' in the repository details area.

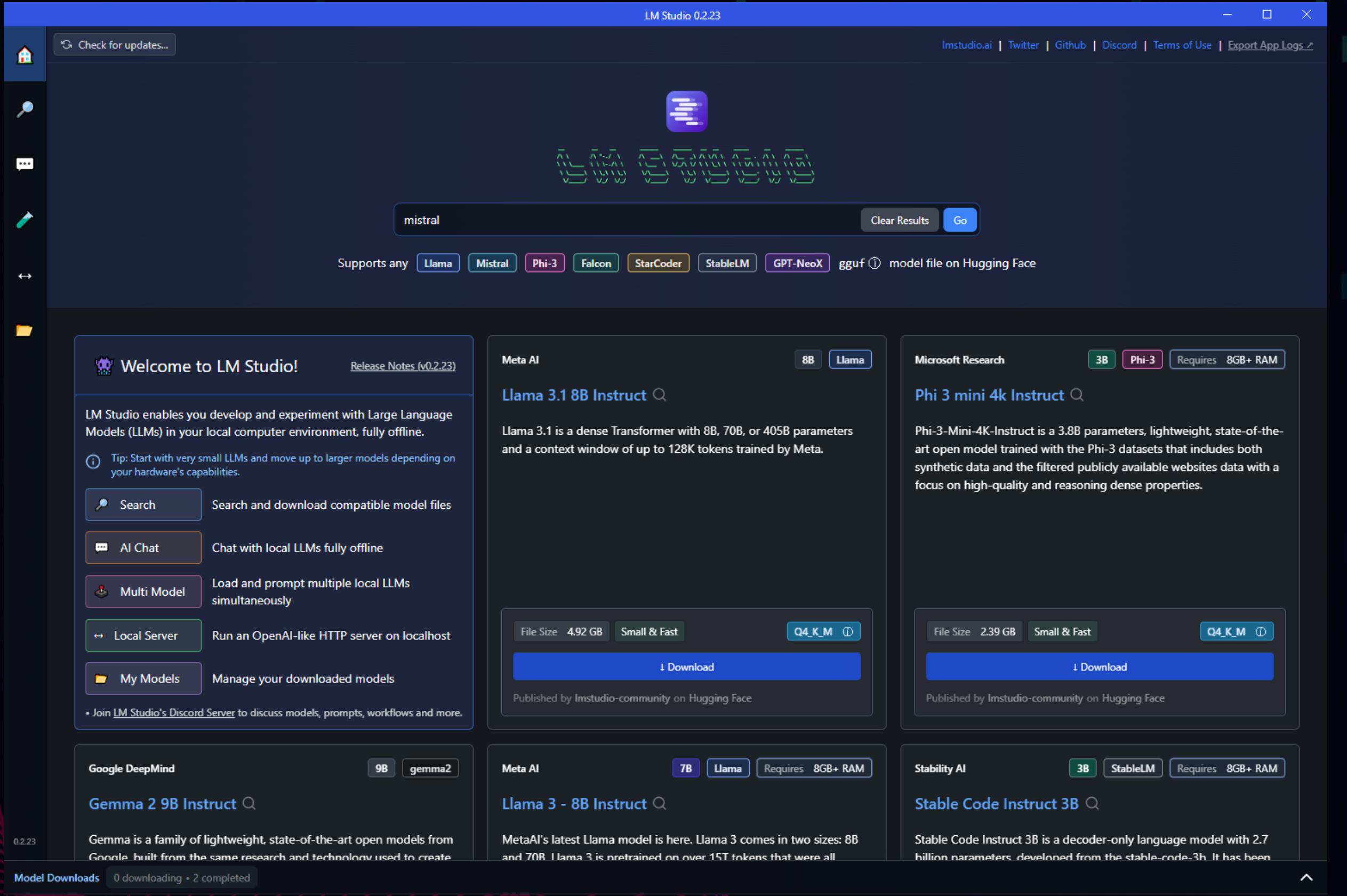
IA Off-Line

Em Servidor ou uma VPS

- A instalação do Ollama deve ser feita normalmente na VPS ou servidor externo/dedicado
- Para dar acesso ao Ollama instalado em nossa VPS, vamos usar a API Ollama4Delphi:
<https://github.com/Code4Delphi/Ollama4Delphi>
- Para instalar: `http://<ip-ou-dominio-do-servidor>:9004/api/chat`



IA Off-Line LM Studio



LM Studio 0.2.23

lmstudio.ai | Twitter | Github | Discord | Terms of Use | Export App Logs

mistral

Clear Results Go

Supports any Llama, Mistral, Phi-3, Falcon, StarCoder, StableLM, GPT-NeoX, gguf model file on Hugging Face

Welcome to LM Studio!

Release Notes (v0.2.23)

LM Studio enables you develop and experiment with Large Language Models (LLMs) in your local computer environment, fully offline.

Tip: Start with very small LLMs and move up to larger models depending on your hardware's capabilities.

Search Search and download compatible model files

AI Chat Chat with local LLMs fully offline

Multi Model Load and prompt multiple local LLMs simultaneously

Local Server Run an OpenAI-like HTTP server on localhost

My Models Manage your downloaded models

Join LM Studio's Discord Server to discuss models, prompts, workflows and more.

Meta AI

Llama 3.1 8B Instruct

8B Llama

File Size 4.92 GB Small & Fast Q4_K_M

↓ Download

Published by lmstudio-community on Hugging Face

Microsoft Research

Phi 3 mini 4k Instruct

3B Phi-3 Requires 8GB+ RAM

File Size 2.39 GB Small & Fast Q4_K_M

↓ Download

Published by lmstudio-community on Hugging Face

Google DeepMind

gemma2

Meta AI

Llama 3 - 8B Instruct

7B Llama Requires 8GB+ RAM

File Size 4.92 GB Small & Fast Q4_K_M

↓ Download

Published by lmstudio-community on Hugging Face

Stability AI

Stable Code Instruct 3B

3B StableLM Requires 8GB+ RAM

File Size 2.39 GB Small & Fast Q4_K_M

↓ Download

Published by lmstudio-community on Hugging Face

0.2.23

Model Downloads 0 downloading • 2 completed



Treinamento de IA

AnythingLLM

AnythingLLM | Your personal LLM trained on anything

v1.5.4

Anything LL M

+ New Workspace

Mistral001

Welcome to AnythingLLM, AnythingLLM is an open-source AI tool by Mintplex Labs that turns anything into a trained chatbot you can query and chat with. AnythingLLM is a BYOK (bring-your-own-keys) software so there is no subscription, fee, or charges for this software outside of the services you want to use with it.

AnythingLLM is the easiest way to put powerful AI products like OpenAI, GPT-4, LangChain, PineconeDB, ChromaDB, and other services together in a neat package with no fuss to increase your productivity by 100x.

AnythingLLM can run totally locally on your machine with little overhead you won't even notice it's there! No GPU needed. Cloud and on-premises installation is available as well. The AI tooling ecosystem gets more powerful everyday. AnythingLLM makes it easy to use.

Create an issue on Github

How do I get started?!

It's simple. All collections are organized into buckets we call "Workspaces". Workspaces are buckets of files, documents, images, PDFs, and other files which will be transformed into something LLM's can understand and use in conversation.

You can add and remove files at anytime.

+ Create your first workspace

Is this like an AI dropbox or something? What about chatting? It is a chatbot isn't it?

AnythingLLM is more than a smarter Dropbox.

AnythingLLM offers two ways of talking with your data:

Query:

Your chats will return data or inferences found with the documents in your workspace it has access to. Adding more documents to the Workspace make it smarter!

AnythingLLM

IA com delphi

PLUGIN
DELPHI AI DEVELOPER

Inspirado no GitHub Copilot

<https://github.com/Code4Delphi/Delphi-AI-Developer>

Chat com interação com banco de dados

The screenshot shows the Delphi AI Developer plugin interface. At the top, there's a window titled "AI DB Chat" with the text "Selecionar todos os clientes que moram na cidade de Ivaipora". Below it, a code editor window titled "Gemini" contains the following SQL query:

```
SELECT * FROM clientes
INNER JOIN cidades ON clientes.id_cidade = cidades.id
WHERE cidades.nome = 'Ivaipora';
```

At the bottom, a MySQL viewer titled "MySQL - code4d_nfce" displays the results of the query:

ID	ID Cidade	Nome	Fantasia	Tipo Juri
1	1	Nome do cliente teste 1	Fantasia do cliente teste 1	J
3	1	Nome do cliente teste 3	Fantasia do cliente teste 3	J
11	1	Nome do cliente teste 11	Fantasia do cliente teste 11	J
12	1	Nome do cliente teste 12	Fantasia do cliente teste 12	F
16	1	Cliente inserido em: 2024-08-26 17:57:18	Cliente inserido em: 2024-08-26 17:57:07	F
18	1	Cliente inserido em: 2024-08-26 17:57:55	Cliente inserido em: 2024-08-26 17:57:07	F
19	1	Cliente inserido em: 2024-08-26 17:58:46	Cliente inserido em: 2024-08-26 17:57:07	J
20	1	Cliente inserido em: 2024-08-26 22:12:01	Cliente inserido em: 2024-08-26 17:57:07	F
21	1	Cliente inserido em: 2024-08-26 22:53:07	Cliente inserido em: 2024-08-26 17:57:07	J
22	1	Cliente inserido em: 2024-08-26 22:53:16	Cliente inserido em: 2024-08-26 17:57:07	F
102	1	Nome do cliente teste 102	Fantasia teste 102	F
567	1	Nome do cliente teste 567	Fantasia teste 567	J
695	1	Nome do cliente teste 695	Fantasia teste 695	J
702	1	Nome do cliente teste 702	Fantasia teste 702	F

IAs On-Line
e Off-Line



Code Completion

```
public
    property Nome: string read FName write FName;
    property Endereco: string read FEndereco write FEndereco;
    property Bairro: string read FBairro write FBairro;
    property Numero: Integer read FNumero write FNumero;
    property Telefone: string read FTelefone write FTelefone;
    property Email: string read FEmail write FEmail;
    property DataNascimento: TDateTime read FDataNascimento write FDataNascimento;
    property Apelido: string read FApelido write FApelido;
    |
    property Cep: Integer read FCep write FCep; ←
```

```
procedure Clear;
end;
```

IA com delphi

Delphi + GitHub Copilot

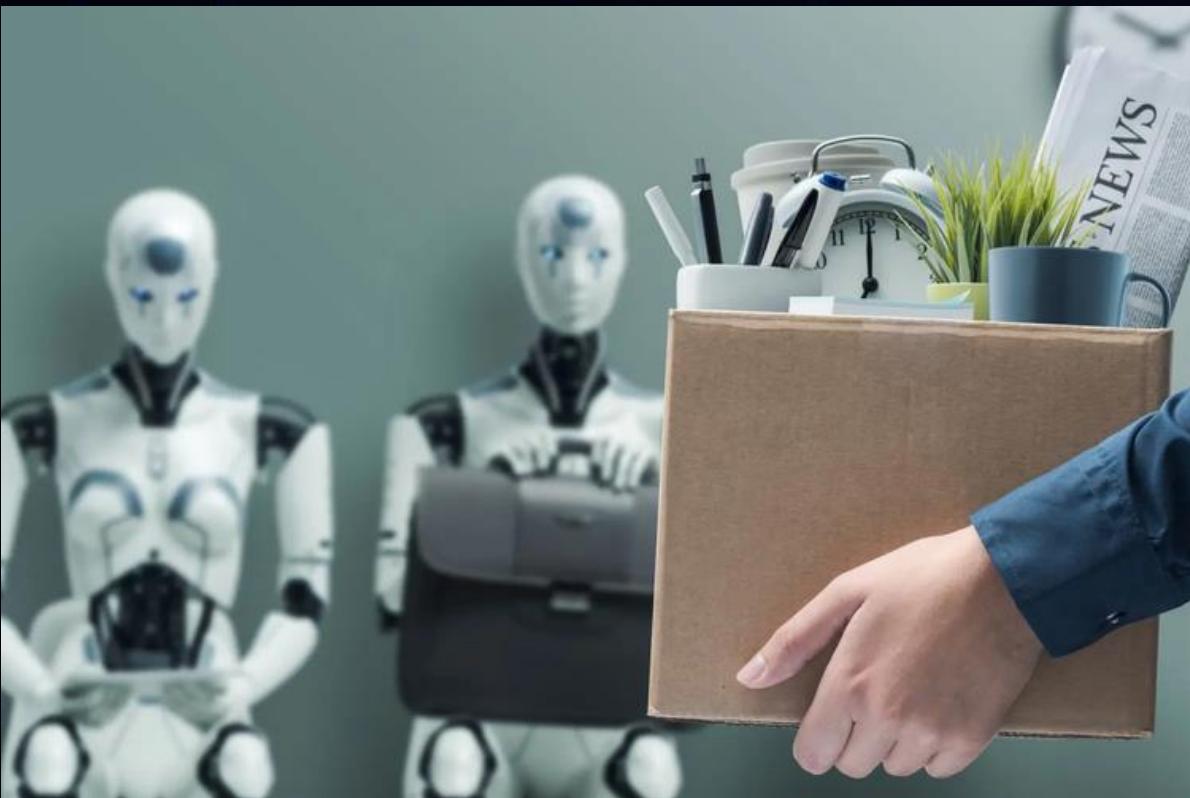
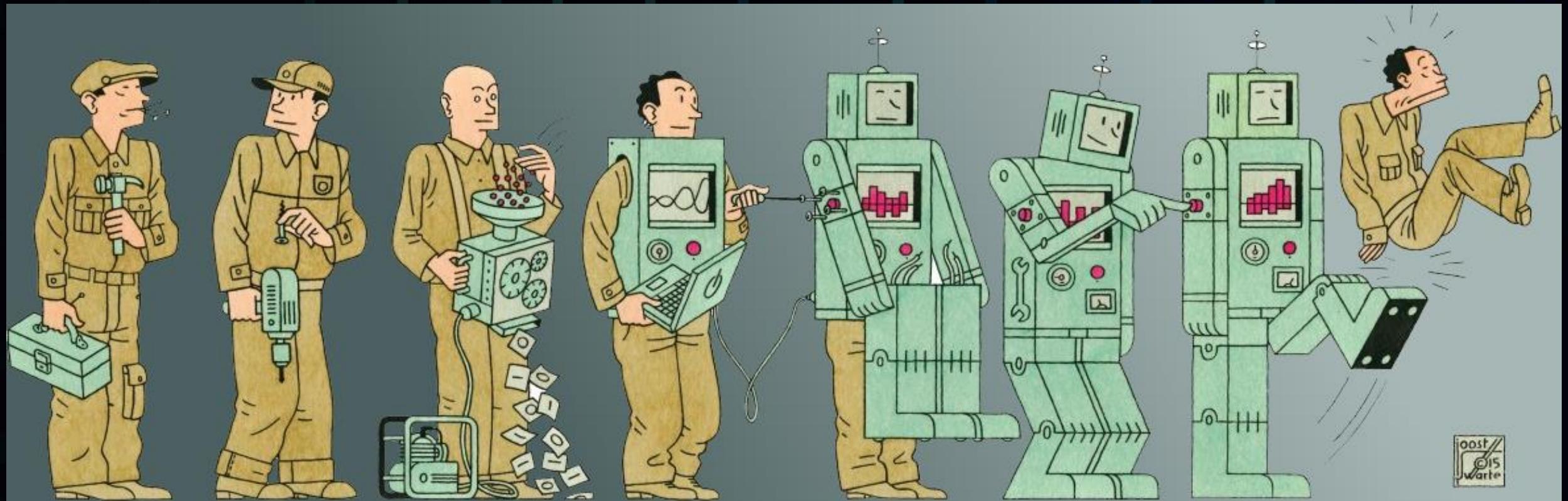
<https://code4delphi.com.br/blog/delphi-github-copilot/>



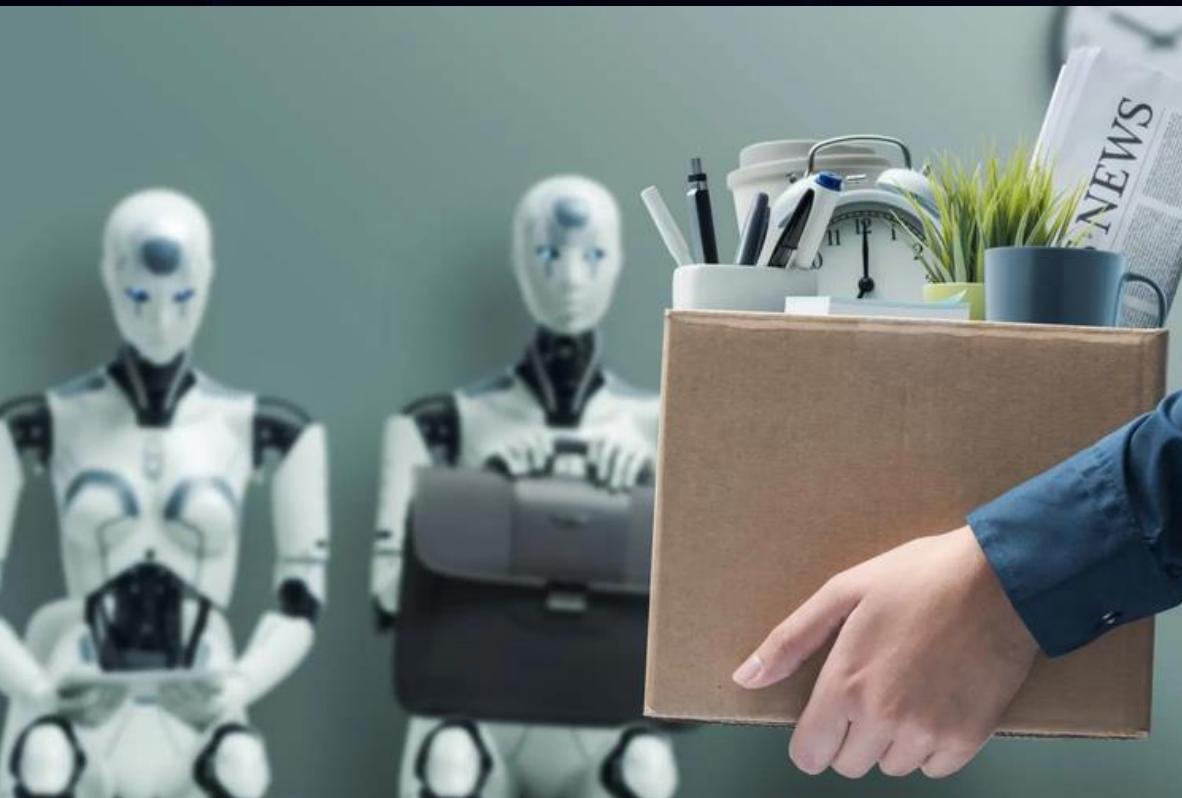
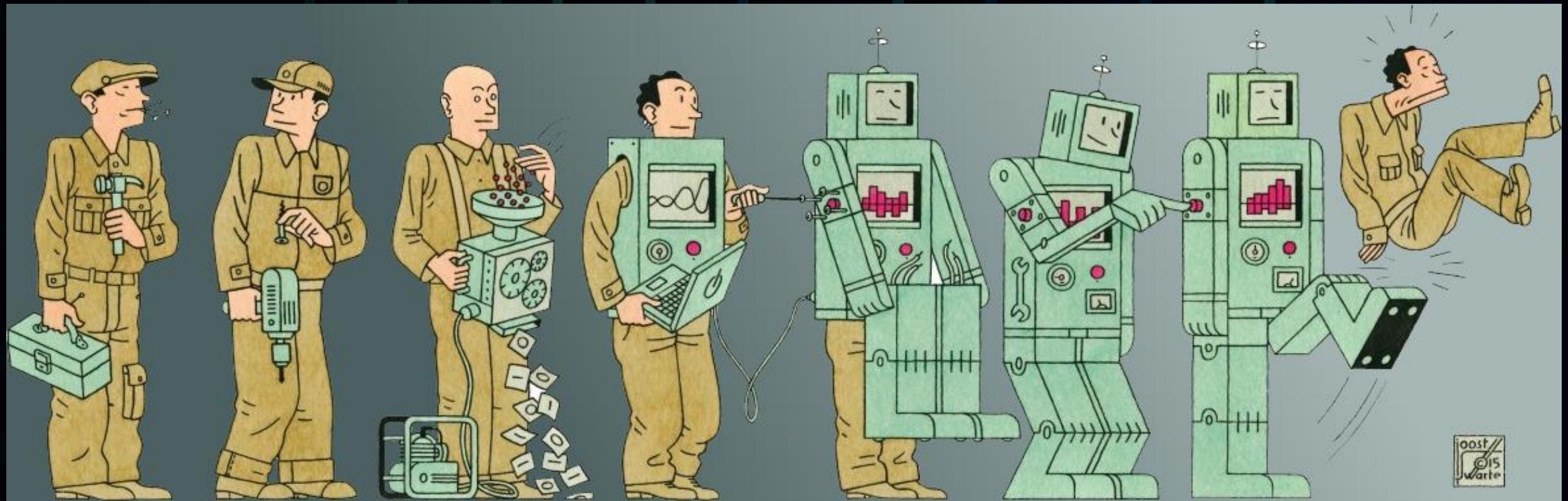
GitHub
Copilot
Delphi



IA vai roubar nossos empregos?



IA vai roubar nossos empregos?

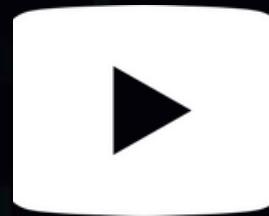


Você não será substituído pela IA. Você será substituído por pessoas que usam a IA melhor do que você.

Contatos



github.com/Code4Delphi



youtube.com/@Code4Delphi



[Code4Delphi](https://www.instagram.com/Code4Delphi)



<https://t.me/Code4Delphi>



contato@code4delphi.com.br



César Cardoso