

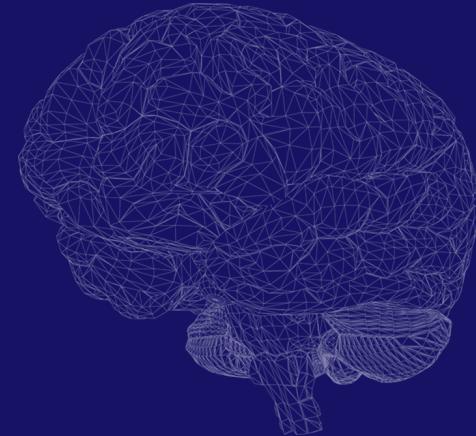




Marcelo Filipak
Diretor de Inovação da AFRAC
CEO Filiart

I.A. NA PRÁTICA

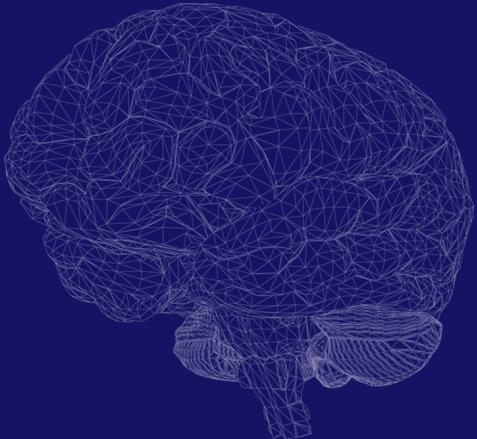
cases em automação comercial



I.A. NA PRÁTICA

cases em automação comercial

Marcelo Filipak, Engenheiro Eletrônico e Cientista de dados.



- . Treino redes neurais há 7 anos.
- . Me especializei em criar modelos.
- . Desenvolvo projetos de modelos de deep learning com aplicações na área de saúde, segurança pública, controle de tráfego, logística, inteligência fiscal, controle aduaneiro, controle de qualidade ...
- . Trabalhei com hardware e software para automação comercial de 1990 a 2018 na BEMATECH e na TOTVS.
- . CEO na FiliArt - Projetos e Advisory em I.A

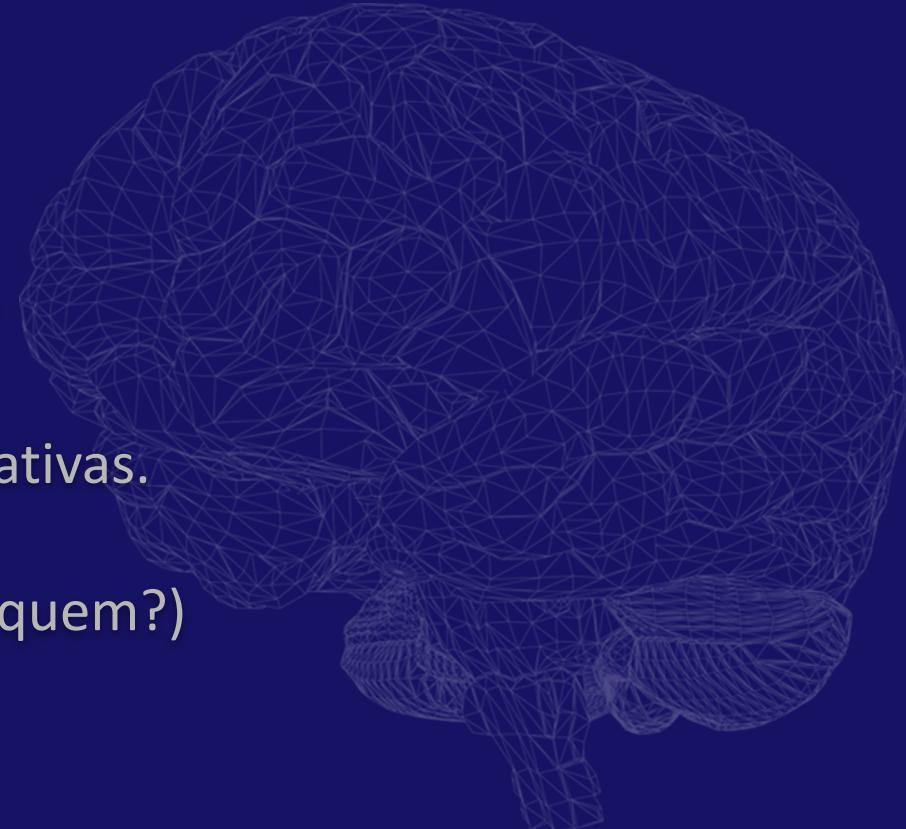
Parte 1 - Cases

Parte 2 - Que tipo de I.A eu preciso?

Parte 3 - Avanços Recentes e Expectativas.

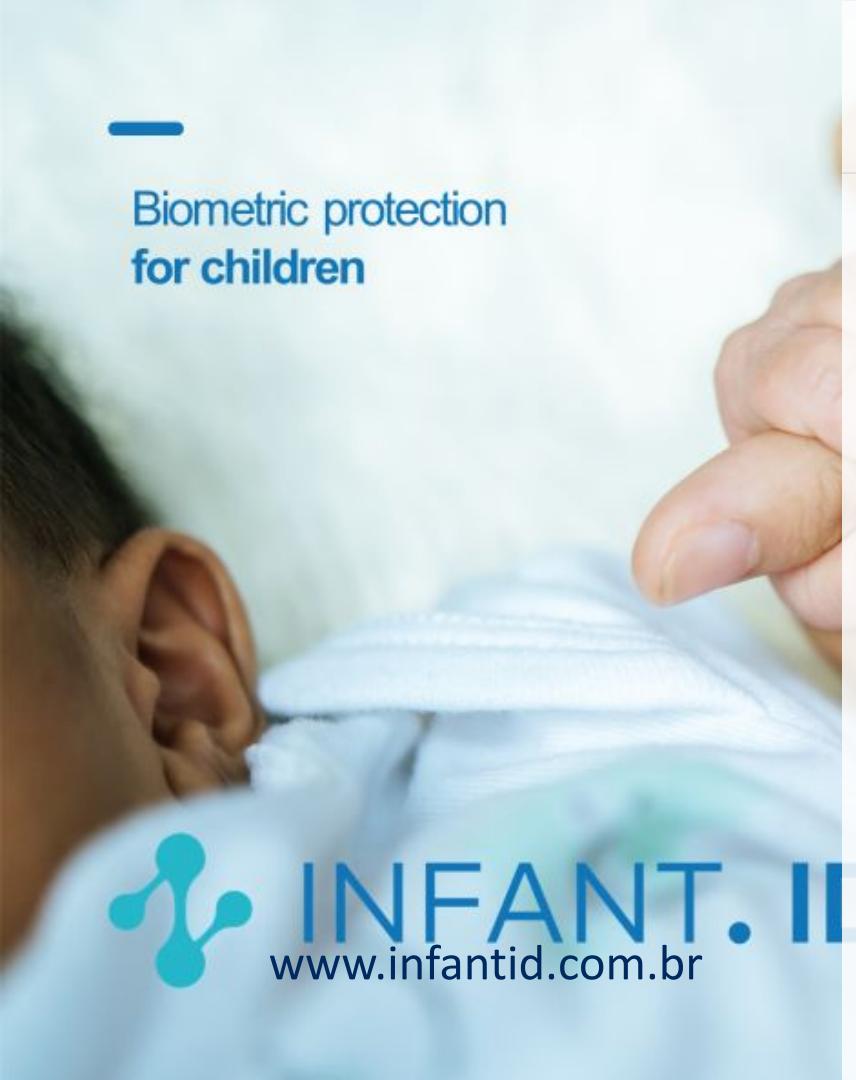
BÔNUS 1 - Evolução da I.A (Quem é quem?)

BÔNUS 2 - Onde e o que aprender.



Biometric protection
for children





Biometric protection
for children

 INFANT. ID
www.infantid.com.br

Google

biometria recém nascidos

Todas Shopping Imagens Vídeos Notícias Web Maps Mais

[www.youtube.com > watch](http://www.youtube.com/watch)

Biometria de recém-nascidos no Piauí é destaque no Jornal ...



[globoplay.globo.com > ...](http://globoplay.globo.com)

Jornal Hoje | Piauí é o primeiro estado a usar a biometria ...



[globoplay.globo.com > ...](http://globoplay.globo.com)

PITV 2ª Edição | Novo RG dos bebês: Piauí é o primeiro ...



[globoplay.globo.com > ...](http://globoplay.globo.com)

Boa Noite Paraná | Começam os testes para a identificação ...



Biometria é coletada dos recém-nascidos e das mães. Programa deve ser ampliado para todas as maternidades públicas do estado.

Globoplay · 11 de fev. de 2022

O Jornal Hoje, da TV Globo, exibiu neste sábado (23) uma reportagem especial sobre o nosso projeto-piloto de Identificação Neonatal no Piauí ...

YouTube · Rafael Fonteles · 24 de nov. de 2024

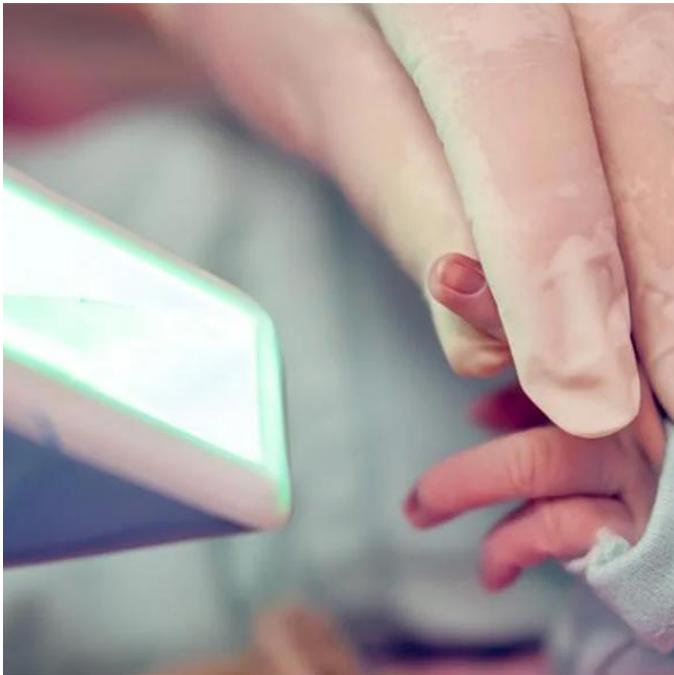
Piauí é o primeiro estado a usar a biometria digital em recém-nascidos - Novidade. 2 min. Piauí é o primeiro estado a usar a biometria digital ...

Globoplay · 23 de nov. de 2024

Novo RG dos bebês: Piauí é o primeiro estado a usar biometria digital em recém-nascidos.

Globoplay · 24 de nov. de 2024

Projeto - INFANT.ID



INFANT.ID

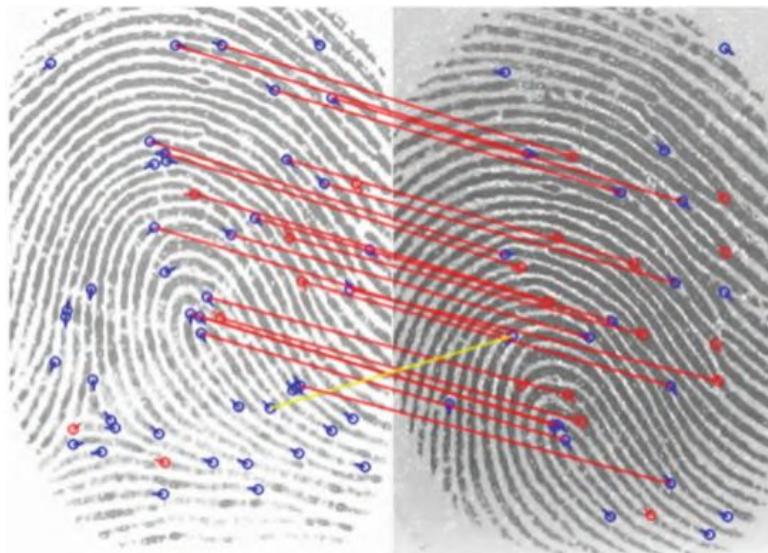
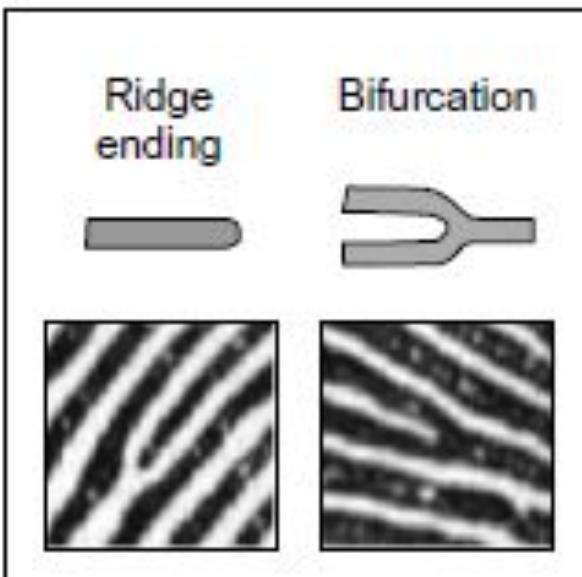
Construindo um mundo mais seguro para as nossas
crianças

[CONHEÇA A PLATAFORMA INFANT.ID](#)



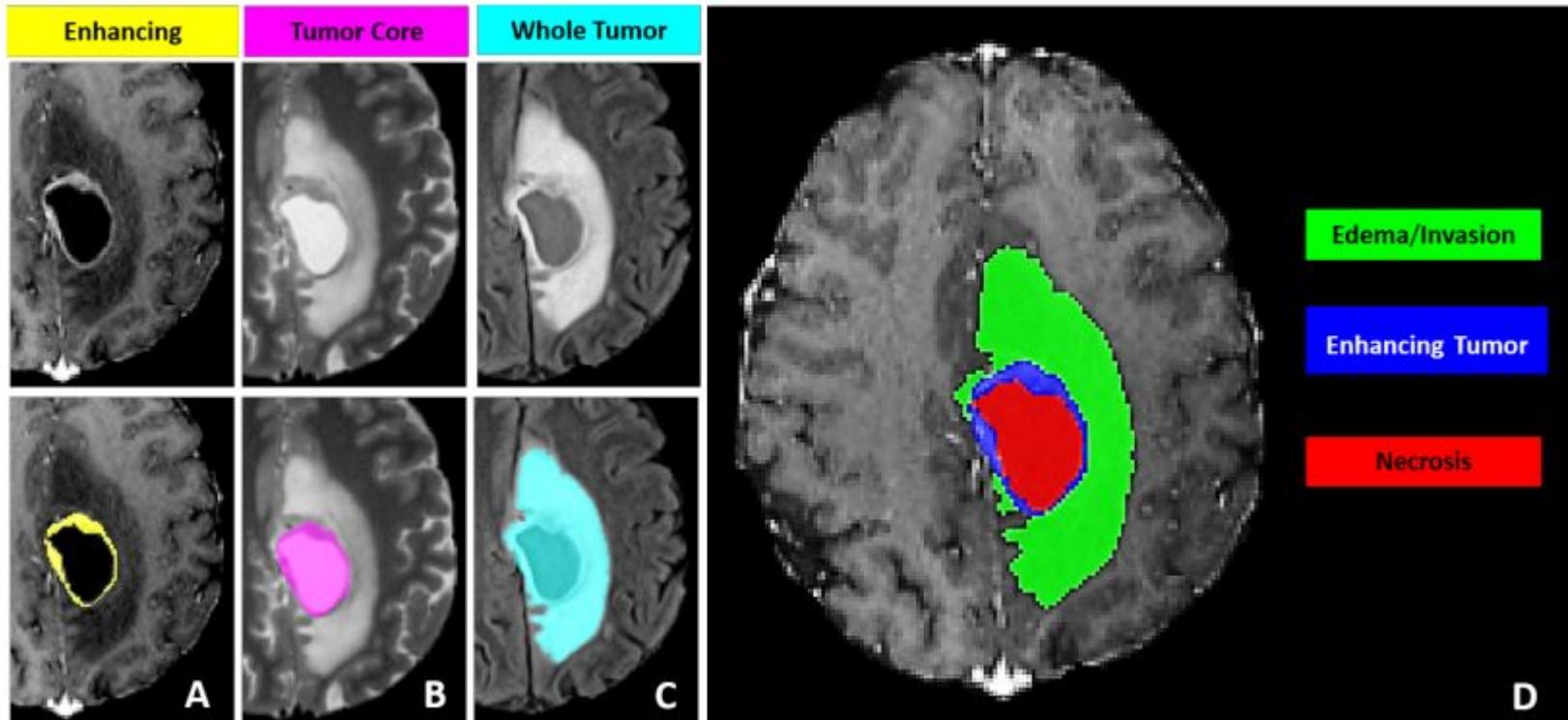


Projeto - INFANT.ID

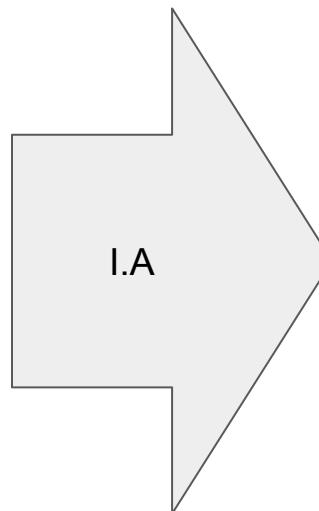
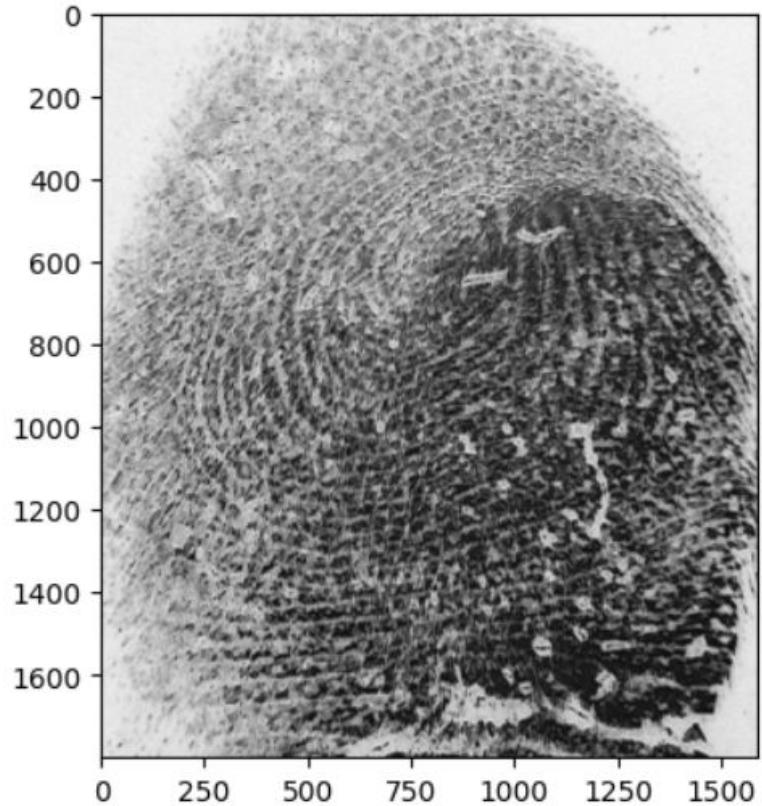


Para adultos é uma tecnologia madura e padronizada.

Problema semelhante



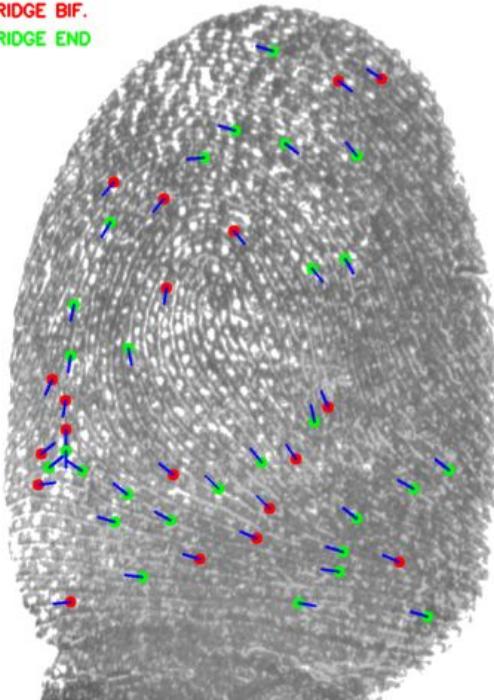
Projeto - INFANT.ID



Projeto - INFANT.ID

MINUTIAE LABEL

- RIDGE BIF.
- RIDGE END



MINUTIAE PREDICTION

- RIDGE BIF.
- RIDGE END

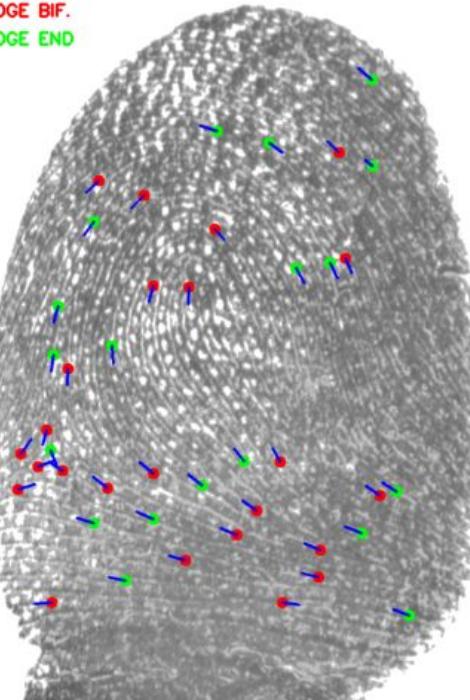
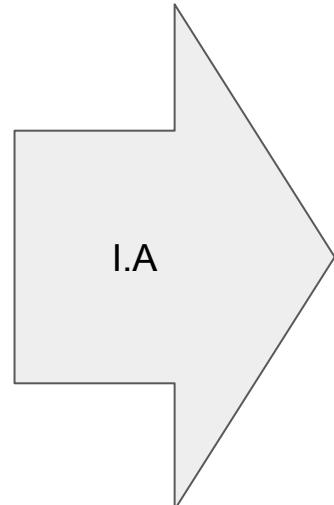


Imagen “boa” de
recém nascido.



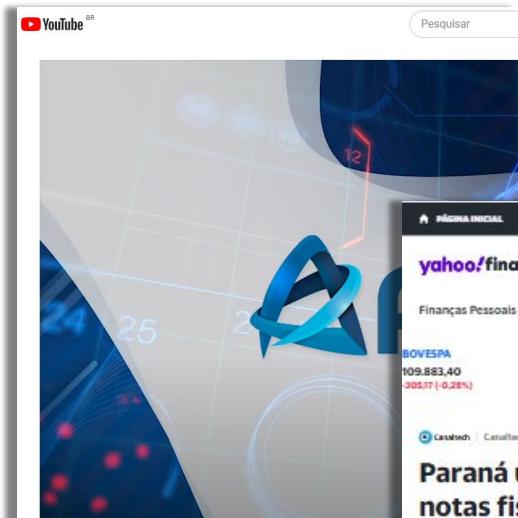
Projeto - INFANT.ID

Treinar redes pode levar tempo e \$

epoch_DICE_VALLEYS
tag: epoch_DICE_VALLEYS



Projeto - AFRAC Nota Certa



yahoo! finanças

PÁGINA INICIAL MAIL NOTÍCIAS FINANÇAS ESPORTES CELEBRIDADES VIDA E ESTILO

Buscar por notícias, códigos de negociação ou empresas

Finanças Pessoais Investimentos Empreendedorismo Trabalho Líderes Agro

BOVESPA 109.883,40 -305,37 (-0,28%) Merval 38.390,84 +233,89 (+0,61%) MXX 51.020,65 +221,19 (+0,44%) PETROLEO CRU 74,30 +0,05 (+0,07%)

Canaltech Canaltech

Paraná usará inteligência artificial para preencher e corrigir notas fiscais

Kaique Lima | 25 de agosto de 2022 às 15h20

A Secretaria da Fazenda do Paraná (SEFA/PR) firmou uma parceria com a Associação Brasileira de Tecnologia para o Comércio e Serviços (AFRAC) para o desenvolvimento de um projeto-piloto que envolve o uso da inteligência artificial para preencher e corrigir notas fiscais. O objetivo do acordo é reduzir a complexidade fiscal para os varejistas e para o fisco.



Paraná usará inteligência artificial para preencher e corrigir notas fiscais

Por Kaique Lima | Editado por Cláudio Yuge | 25 de Agosto de 2022 às 15h20

compartilhar



A Secretaria da Fazenda do Paraná (SEFA/PR) firmou uma parceria com a Associação Brasileira de Tecnologia para o Comércio e Serviços (AFRAC) para o desenvolvimento de um projeto-piloto que envolve o uso da inteligência artificial para preencher e corrigir notas fiscais. O objetivo do acordo é reduzir a complexidade fiscal para os varejistas e para o fisco.

Comer carne romaneada a ouro faz mal?

1. Bear market do Bitcoin está nos "estágios finais", diz analista da Bloomberg

2. Petrobras reduz diesel e gasolina nas refinarias após meses sem reajustes

Projeto - AFRAC Nota Certa



APP AFRAC



WEB APP HUGGINGFACE



API - WS



Aplicativo AFRAAC



Ao acessar a tela inicial do Aplicativo, busque pela Opção Nota Certa



O aplicativo é
grátis e está
disponível
mesmo para não
associados.

Aplicativo AFRAC

Digite Livremente a
Descrição do
Produto que deseja
descobrir o NCM.



Aplicativo AFRAC



O Aplicativo irá informar uma lista de prováveis NCMs indicando o grau de confiança da resposta.

Reparem que eu propositalmente escrevi “Brhma” errado, então o Nível de confiança foi apenas de 48%

Aplicativo AFRAC

Quando a descrição é um pouco mais completa, os resultados são ainda melhores.



Desta vez o nível de confiança na resposta é de 95%



Baixe o APP e teste com seus próprios produtos

WEB APP - HUGGINGFACE

AFRAC NOTA CERTA

Descrição
Feijão Preto

Clear Submit

NCMs

07133319

07133319	33%
07133399	14%
07133329	5%
20055100	5%
21069090	4%

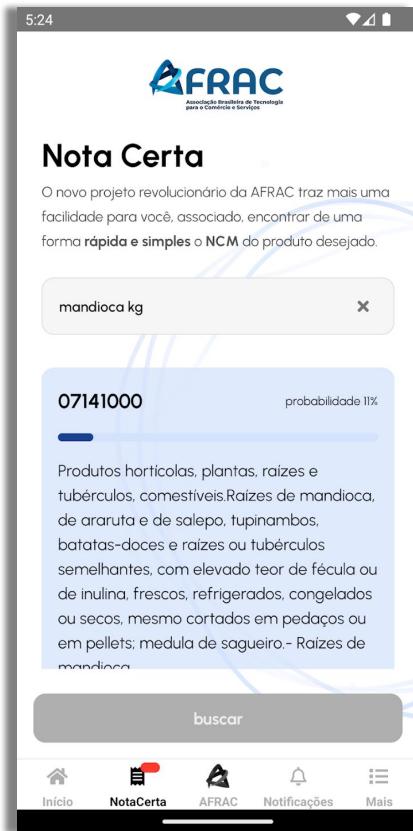
Descrição do NCM

Produtos hortícolas, plantas, raízes e tubérculos, comestíveis. Legumes de vagem, secos, em grão, mesmo pelados ou partidos. - Feijões (<i>Vigna</i> spp., <i>Phaseolus</i> spp.) -- Feijão comum (<i>Phaseolus vulgaris</i>) Preto Outros

https://huggingface.co/spaces/AFRAC/NCM_DEMO
https://github.com/mfilipak/AFRAC_IA



API - AFRAC Nota Certa



```
import requests

url = "https://afrac-ncm-demo.hf.space/api/predict"

payload = {
    "data": ["Coca-Cola PET 21"]
}

headers = {
    "Content-Type": "application/json"
}

response = requests.post(url, json=payload, headers=headers)
result = response.json()

print(result)
```



API - AFRAC Nota Certa



```
import requests

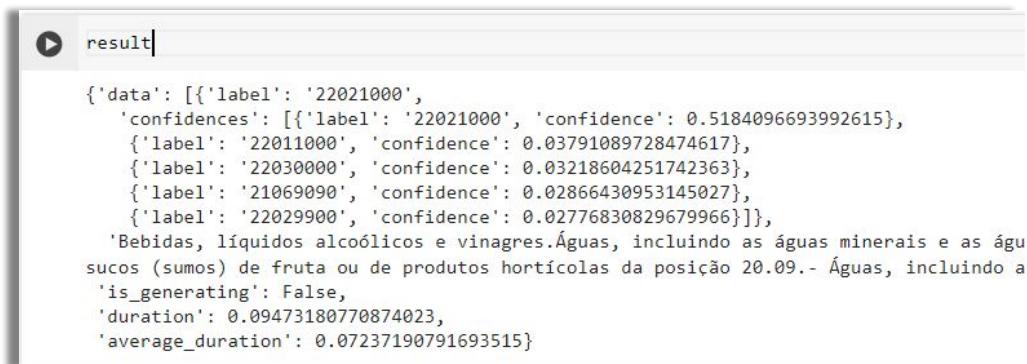
url = "https://afrac-ncm-demo.hf.space/api/predict"

payload = {
    "data": ["Coca-Cola PET 21"]
}

headers = {
    "Content-Type": "application/json"
}

response = requests.post(url, json=payload,
headers=headers)
result = response.json()

print(result)
```



```
result
{
  "data": [
    {
      "label": "22021000",
      "confidences": [
        {
          "label": "22021000",
          "confidence": 0.5184096693992615
        },
        {
          "label": "22011000",
          "confidence": 0.03791089728474617
        },
        {
          "label": "22030000",
          "confidence": 0.03218604251742363
        },
        {
          "label": "21069000",
          "confidence": 0.02866430953145027
        },
        {
          "label": "22029900",
          "confidence": 0.02776830829679966
        }
      ]
    }
  ],
  "text": "Bebidas, líquidos alcoólicos e vinagres. Águas, incluindo as águas minerais e as águas (sumos) de fruta ou de produtos hortícolas da posição 20.09.- Águas, incluindo a",
  "is_generating": false,
  "duration": 0.09473180770874023,
  "average_duration": 0.07237190791693515
}
```

GT IOT&AI AFRAC



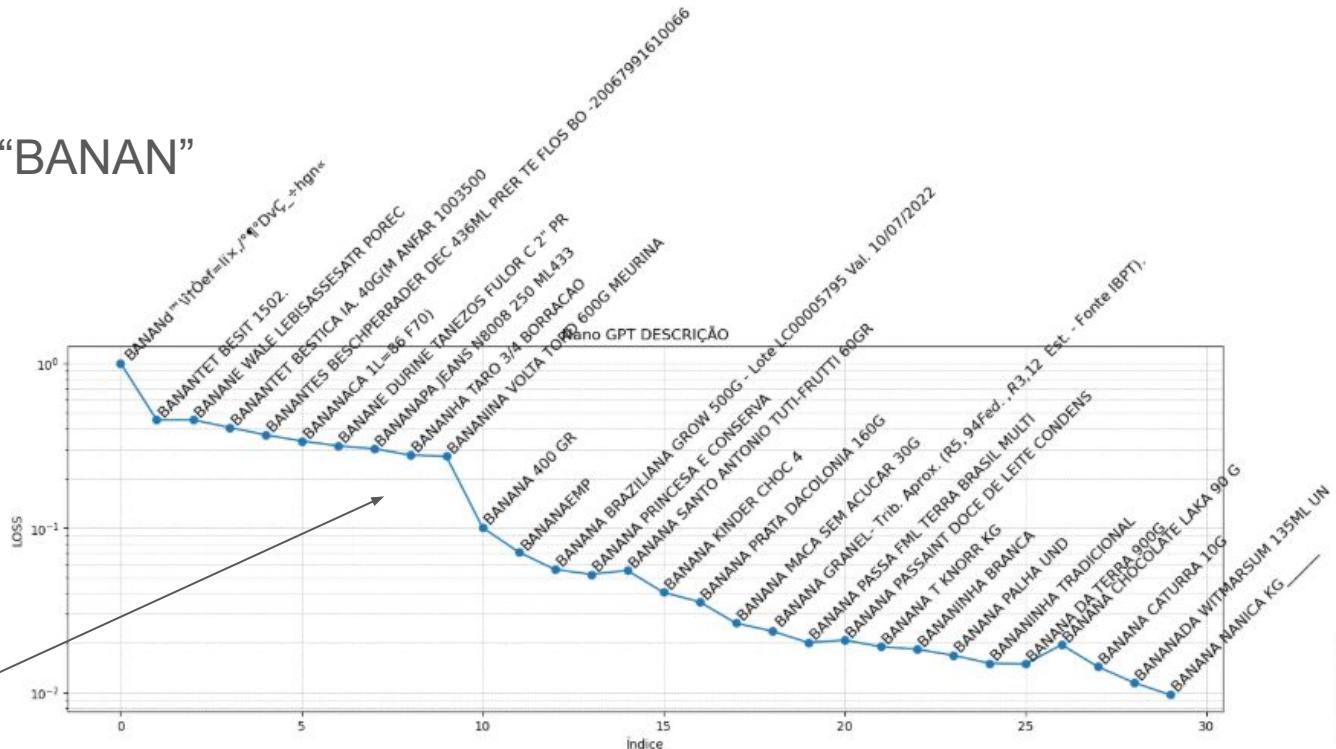
PREDIÇÃO/VERIFICAÇÃO DE GTIN

HIPÓTESE:

- ✓ 1 - Se encontrarmos uma boa representação da descrição.
- ? 2 - Se criarmos uma rede que aprenda a alinhar descrições diferentes do mesmo GTIN.
- ? 3 - Teremos como vetorizar uma tabela e encontrar o GTIN mais próximo mesmo que a descrição seja diferente.

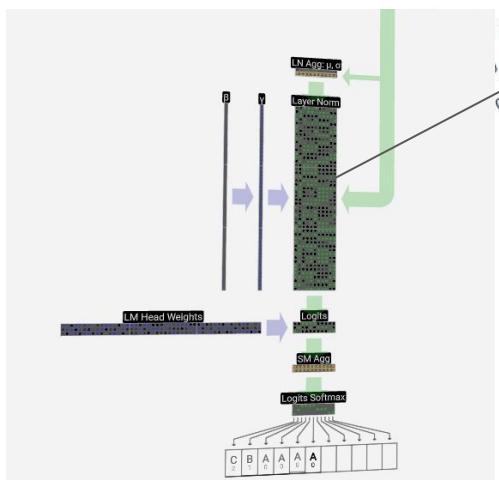
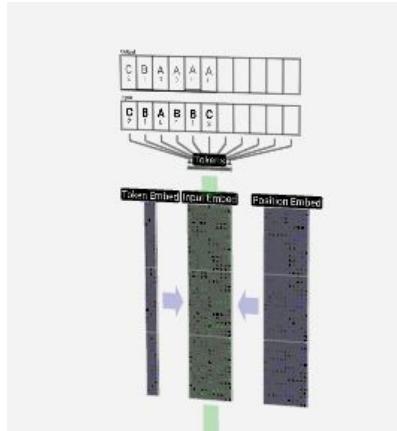
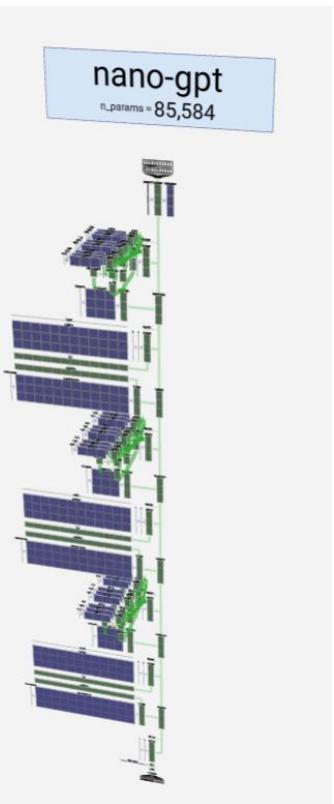
MODELO GENERATIVO DE DESCRIÇÕES - NANOGPT

GT IOT&AI AFRAC

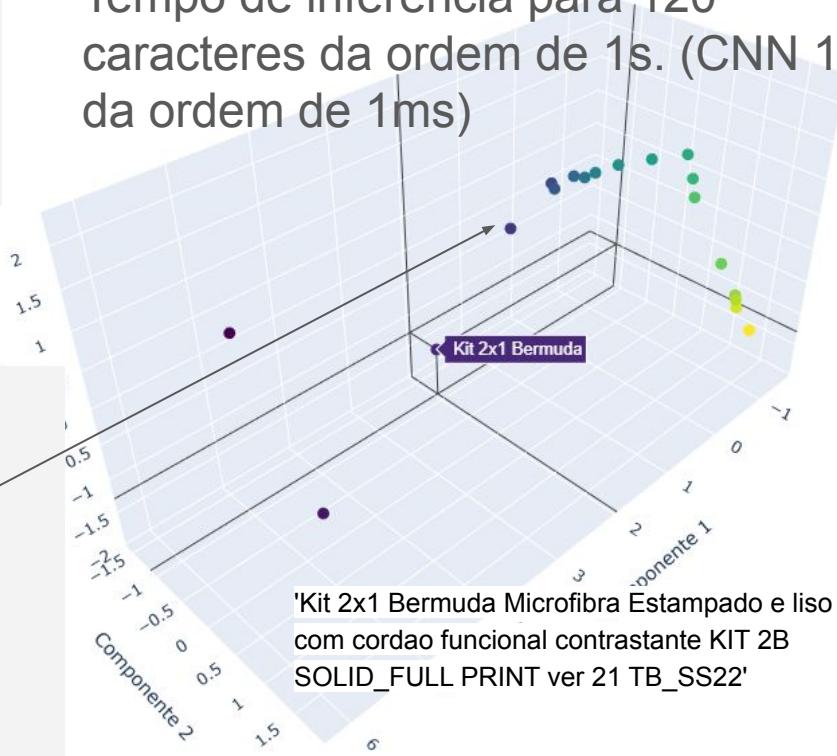


Aproximadamente 21 dias.
(51M de parâmetros)

GT IOT&AI AFRAC



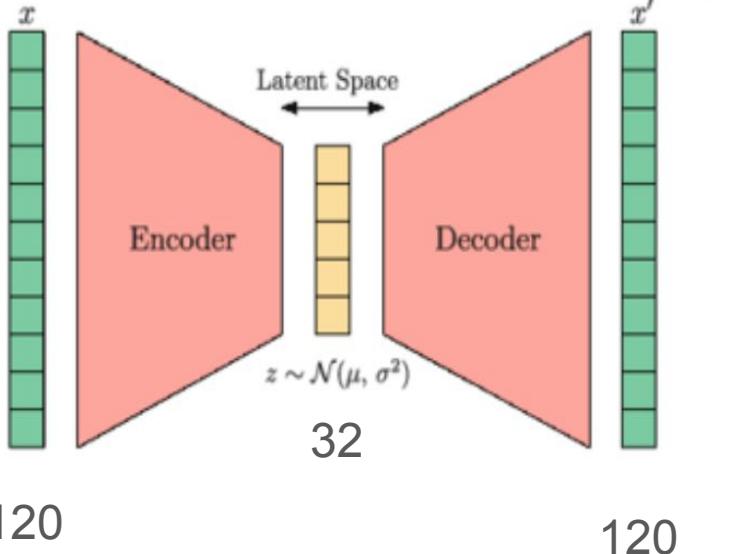
Tempo de inferência para 120 caracteres da ordem de 1s. (CNN 1D da ordem de 1ms)



MODELO GENERATIVO DE DESCRIÇÕES - VAE

BANANA

Input Sample



Aproximadamente 3 dias.
(9M de parâmetros)

GT IOT&AI AFRAC

```
[15]: #s1 = "Caixa Cabo De Rede Utp 4 Pares Cat5e Megatron 305 Homologado"
s1 = "CEBOLITOS"
s2 = "CEBOLA ROXA KG"

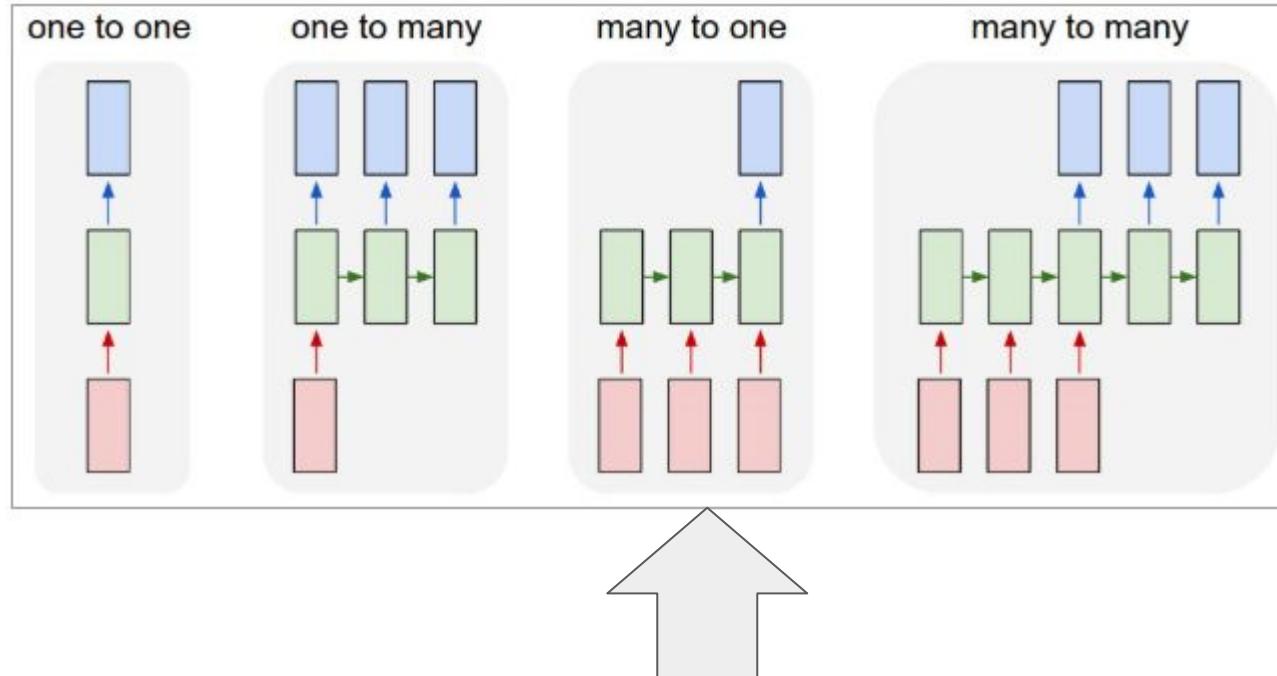
XX = np.array([encode(_) + encode(" ") * (256 - len(_)) for _ in [s1, s2]])
yy = decoder.predict(encoder.predict(XX, verbose=False)[0], verbose=False)

for t in np.arange(-0.2, 1.3, 0.1):
    enc = encoder.predict(XX, verbose=False)[0]
    dec = decoder.predict((t * enc[1] + (1 - t) * enc[0])[None, ...], verbose=False)[0]
    print(f"{t:.2f} {decode(np.argmax(dec, axis=-1))[:60]}")

-0.20 CEBOLYTAS
-0.10 CEBOLITAS
0.00 CEBOLITOS
0.10 CEBOLITES
0.20 CEBOLETES
0.30 CEBOLERI::
0.40 CEBOLER.TR.
0.50 CEBOLOR TTe "
0.60 CEBOLOR ETE "
0.70 CEBOLARAENO "X
0.80 CEBOLAUIESO "G
0.90 CEBOLA YOSA KG
1.00 CEBOLA ROXA KG
1.10 CEBOLA ROJA KG
1.20 CEBOLA TOLA KG.
```

MODELO GENERATIVO DE DESCRIÇÕES - GRU

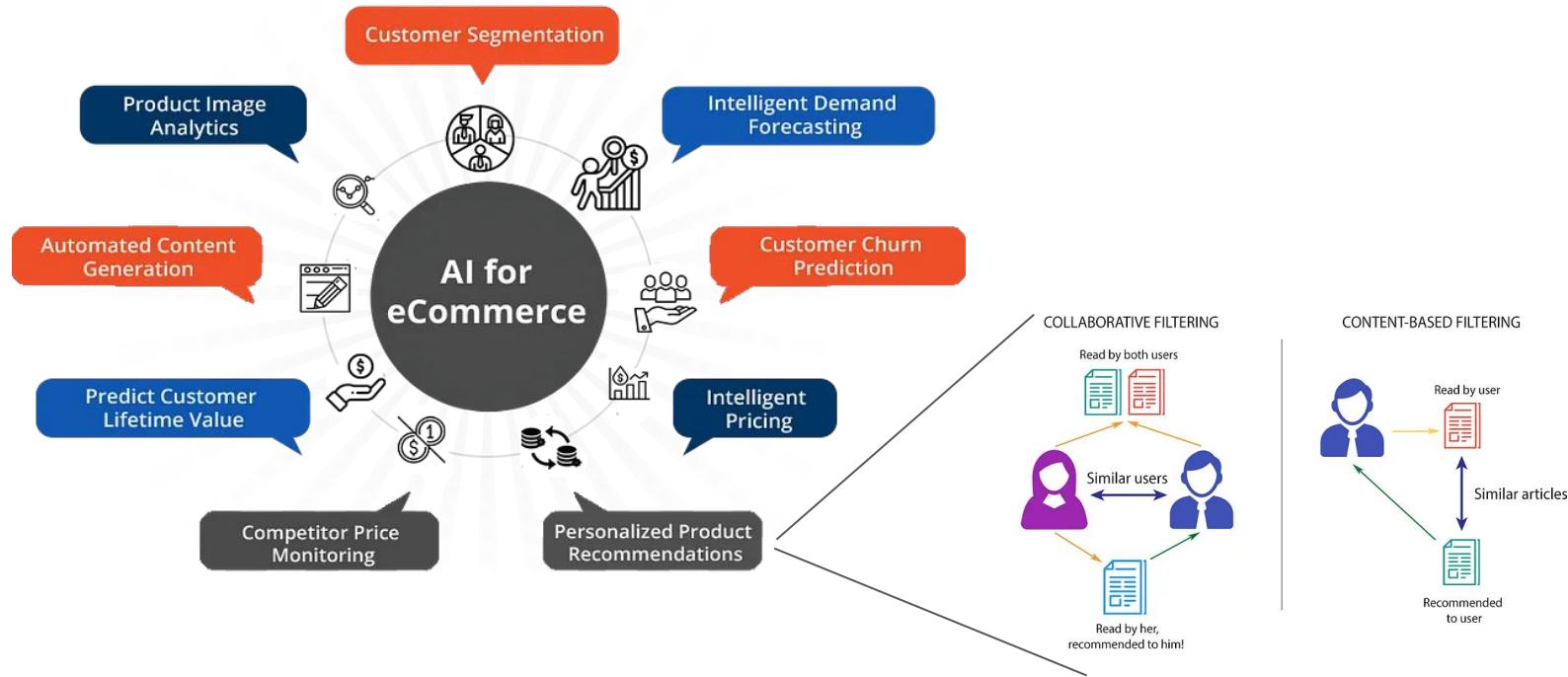
GT IOT&AI AFRAC



Daqui se extrai uma boa representação sem toda a complexidade das arquiteturas de LLM. Possivelmente GRU - Gated Recurrent Unit

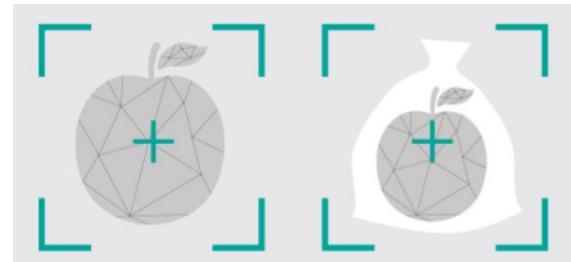
I.A. Discriminativa

I.A. no e-commerce



I.A. Discriminativa

Balanças com I.A.



I.A. Generativa

CHAT-BOTS

Coletando dados
pela interação.

Extender o
suporte para 24x7

Recomendação de
produtos,
“upselling”

Suporte aos
vendedores.

Facilitar tracking dos
pedidos.

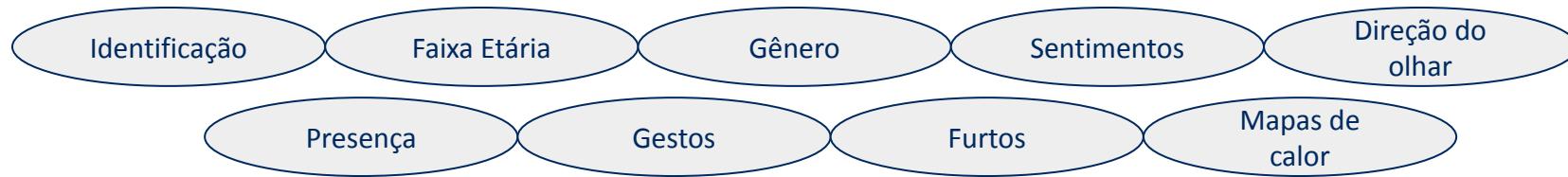
Organizar FAQ e
perguntas
relacionadas.
perplexity.ai



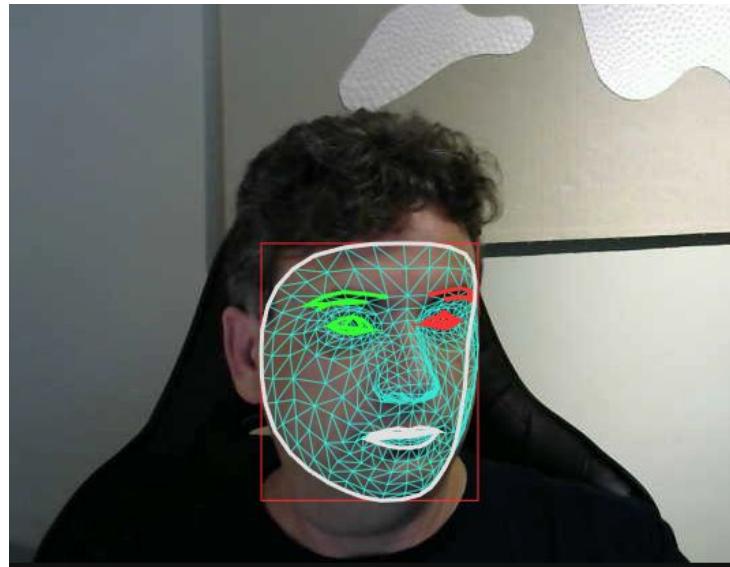
Encontrar uma loja
física.

Coletar feedbacks.

Modelos faciais e corporais



<https://bit.ly/3Y3kvjX>



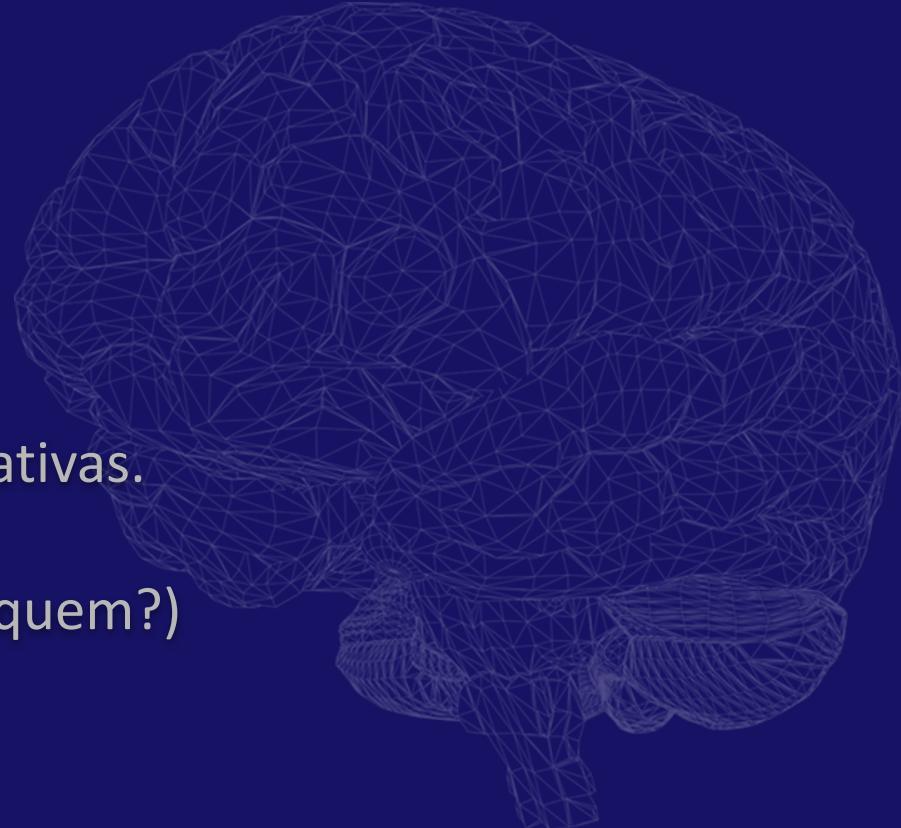
~~Parte 1 - Cases~~

Parte 2 - Que tipo de I.A eu preciso?

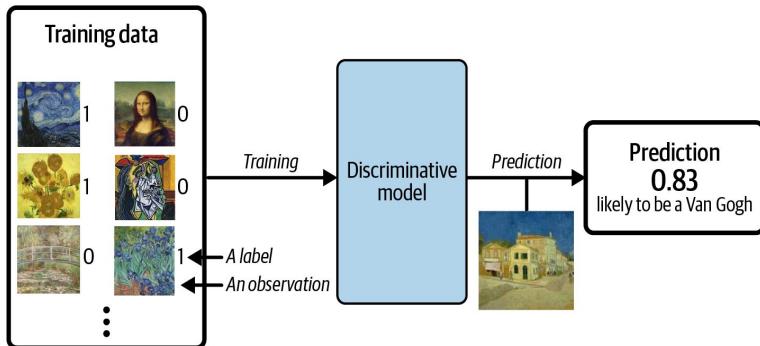
Parte 3 - Avanços Recentes e Expectativas.

BÔNUS 1 - Evolução da I.A (Quem é quem?)

BÔNUS 2 - Onde e o que aprender.

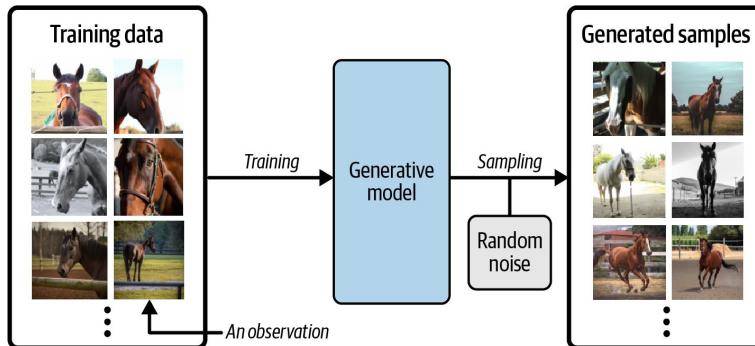


I.A. Discriminativa



Classificação, Regressão
Reconhecimento de imagens
Emoções em texto
Reconhecimento Facial
Contagem de objetos
...

I.A. Generativa



Tradução
Dublagem
Geração de imagens
Chatbots (chat-gpt)
Geração de código
...

São especializadas em problemas diferentes, embora sua arquitetura interna tenha mais semelhanças que diferenças.

Que tipo de modelo eu preciso?

Shallow Learning (Ex: Random Forest)

- . Classificação e Regressão
- Simplificada:** COLLABORATIVE FILTERING, recomendação de vendas, diagnóstico de doenças.
Read by both users.
- . Análise de Crédito: Avaliação de risco de crédito.
- . Detecção de Fraudes: Identificação de transações fraudulentas.
- . Marketing: Segmentação de clientes. Read by her, recommended to him!

Deep Learning



Generative AI

- . Criação de Conteúdo: Geração de texto, imagens, áudio, etc. Extender o suporte para 24x7
- . Design e Arte: Criação de designs gráficos, arte digital.
- . Modelagem e Simulação: Simulação de cenários, modelagem de personagens em jogos.
- . Assistência Cognitiva: Ferramentas de apoio à escrita, composição musical.

Uma indicação e uma contra-indicação.

Shallow Learning (Ex: Random Forest)

Generaliza bem mesmo com poucos dados.

Incapaz de tratar muitos problemas complexos.

Deep Learning

Capaz de tratar problemas complexos.

Se treinada com poucos dados tem overfit (generaliza mal)

Generative AI

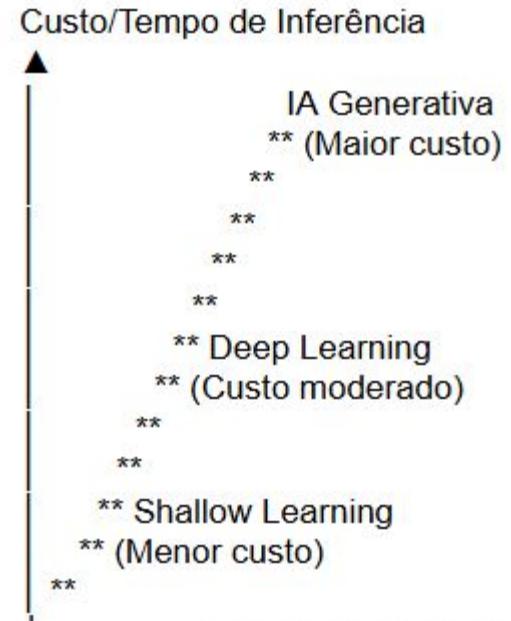
Criativa, tem uma compreensão ampla.

Alucina, precisa ser “controlada”.

Trade-off



Preciso de uma figura que mostre que indo de shallow learning para deep learning e depois para I.A generativa temos de um lado um aumento no custo e no tempo de inferência e no outro uma menor capacidade de problemas complexos



Principais características do trade-off:

1. Shallow Learning

- Custo: Requer ~1-100 GPU horas para treinamento [1](#) [3](#)
- Inferência: <10ms por predição em CPUs comuns [3](#) [6](#)
- Limitações: Eficaz apenas para padrões lineares ou de baixa dimensionalidade [4](#) [6](#)

2. Deep Learning

- Custo: 1,000-100,000 GPU horas (ex: AlphaGo custou milhões) [1](#) [3](#)
- Inferência: 50-500ms usando GPUs especializadas [6](#) [12](#)
- Vantagem: Resolve problemas não-lineares complexos como reconhecimento de imagens [1](#) [6](#)

3. IA Generativa

- Custo: >1M GPU horas (ex: GPT-3 custou ~\$12 milhões) [1](#) [3](#)
- Inferência: 500ms-5s mesmo com hardware otimizado [3](#) [6](#)
- Risco: Tendência a overfitting e alto consumo energético [1](#) [6](#)

Paradoxo da complexidade:

Embora modelos mais profundos tenham **capacidade teórica** superior ($\approx 97\%$ de acurácia em ImageNet) [1](#), na prática enfrentam:

1. **Lei de rendimentos decrescentes**

Melhorias de 0.1% na acurácia podem exigir 10x mais recursos [1](#) [3](#)

2. **Custo ambiental**

Treinar um modelo grande emite ≈ 284 ton CO₂ equivalente [1](#)

3. **Complexidade vs Generalização**

Modelos muito profundos (>100 layers) mostram degradação de performance [6](#) [14](#)

Para aplicações práticas, estudos recomendam:

```
python
```

```
if latency < 100ms and budget < $10k:  
    use shallow models  
elif 100ms < latency < 1s and budget < $1M:  
    use deep learning  
else:  
    consider generative AI  
```[3][6][12]
```



# Modelos prontos - onde?

HUGGING FACE

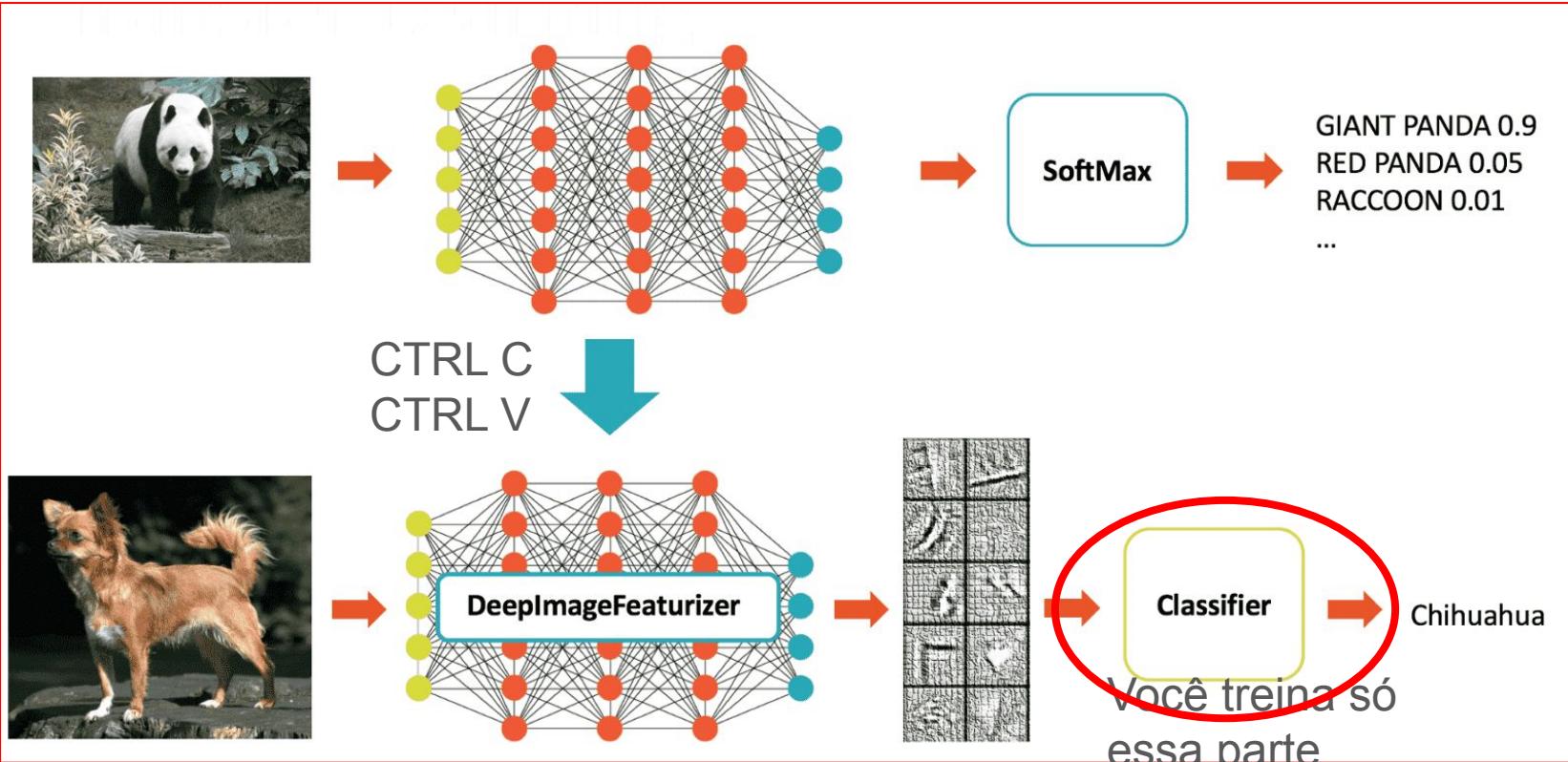


**1** - Procure um modelo pronto para o seu problema (alguns como detecção de face, biometria facial, landscapes faciais, detecção de objetos, ...) são treinados em big data e seguramente vão ser melhores que treinar do zero.

**2** - Se ele estiver no caminho, é possível fazer um “fine tuning” ou “transfer learning”. Aqui você pode congelar parte do modelo e treinar somente uma parte pequena e com relativamente poucos dados.

**3** - Decida se está bom o suficiente ou se você vai treinar um modelo próprio. Nesse cenário você vai precisar obter, tratar ou construir dados e prepará-los para treinar modelos.

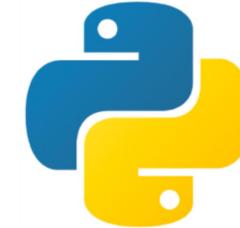
# “Transfer learning” pode poupar muito tempo e \$



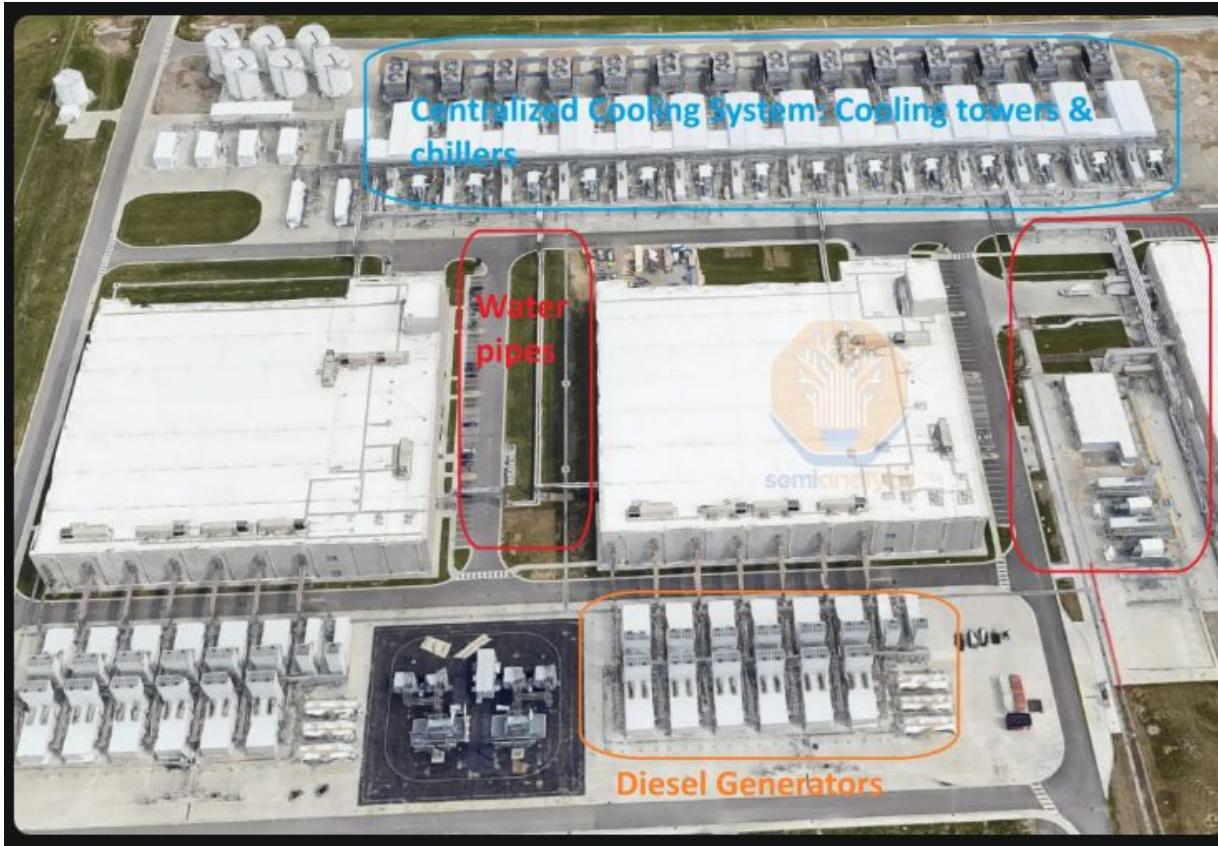
**Que tipo de ambiente eu preciso? (para treinar modelos “simples” ou finetunning / inferência)**



Google  
colab



Que tipo de ambiente eu preciso?  
(para treinar LLMs do zero. ou para inferência em escala)



# Que tipo de ambiente eu preciso? para rodar meus modelos/aplicação



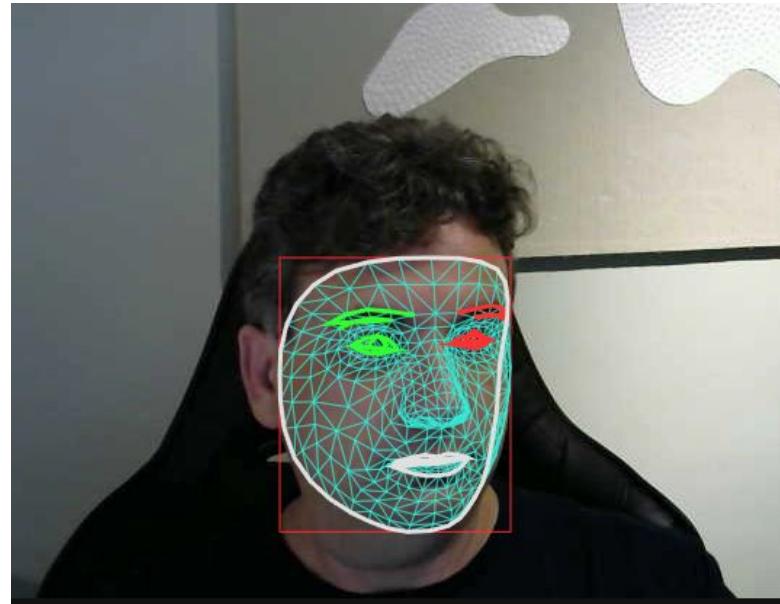
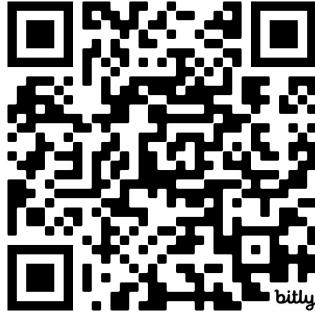
Local ou Nuvem?  
Depende da aplicação.

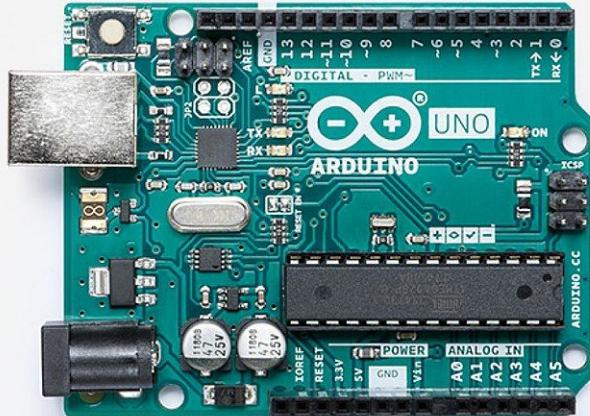


# Rode agora no seu smartphone.

- . O mais comum é rodar em web services (facilidade de prototipação)
- . Muitos modelos complexos rodam até num smartphone. (Exceção: LLMs)
- . Mesmo LLMs podem rodar num desktop.

<https://bit.ly/3Y3kvjX>

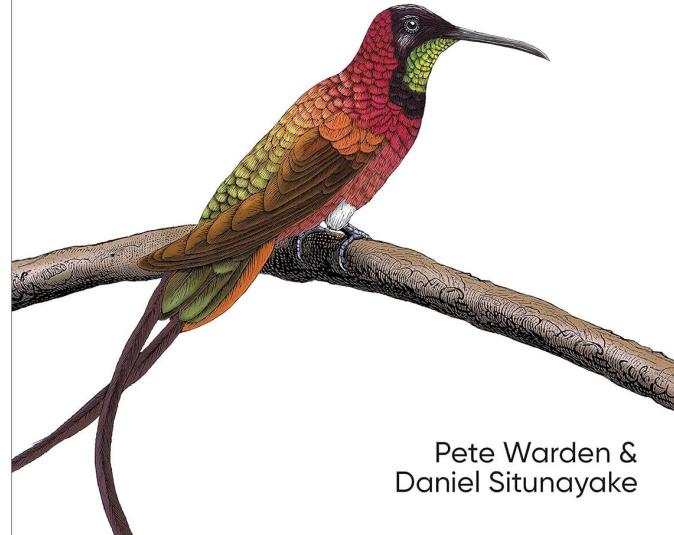




O'REILLY®

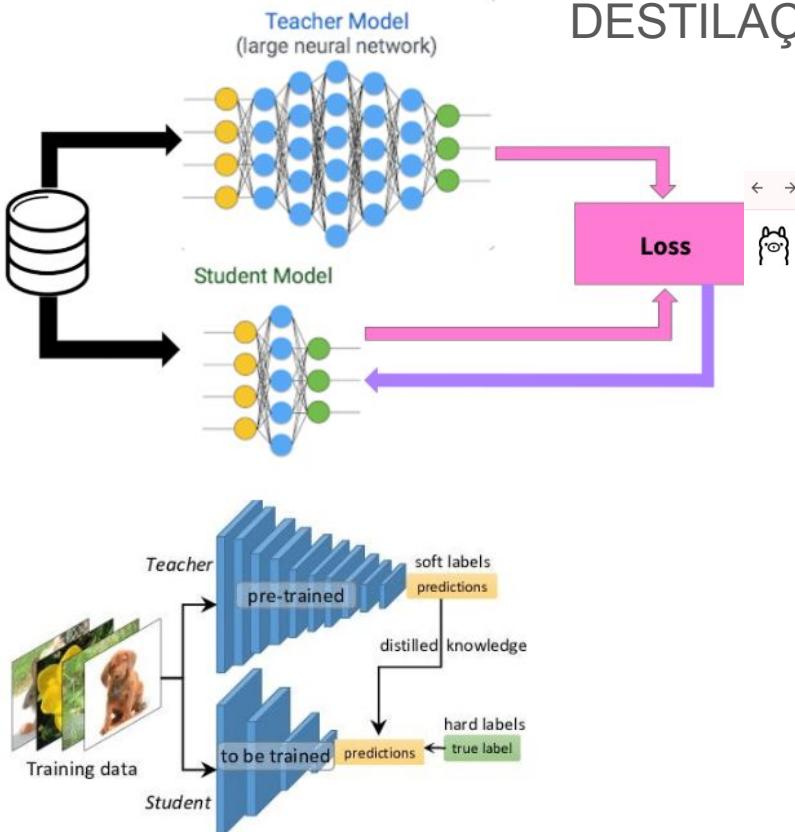
# TinyML

Machine Learning with TensorFlow Lite on  
Arduino and Ultra-Low Power Microcontrollers



Pete Warden &  
Daniel Situnayake

# DESTILAÇÃO / SMALL LLMs

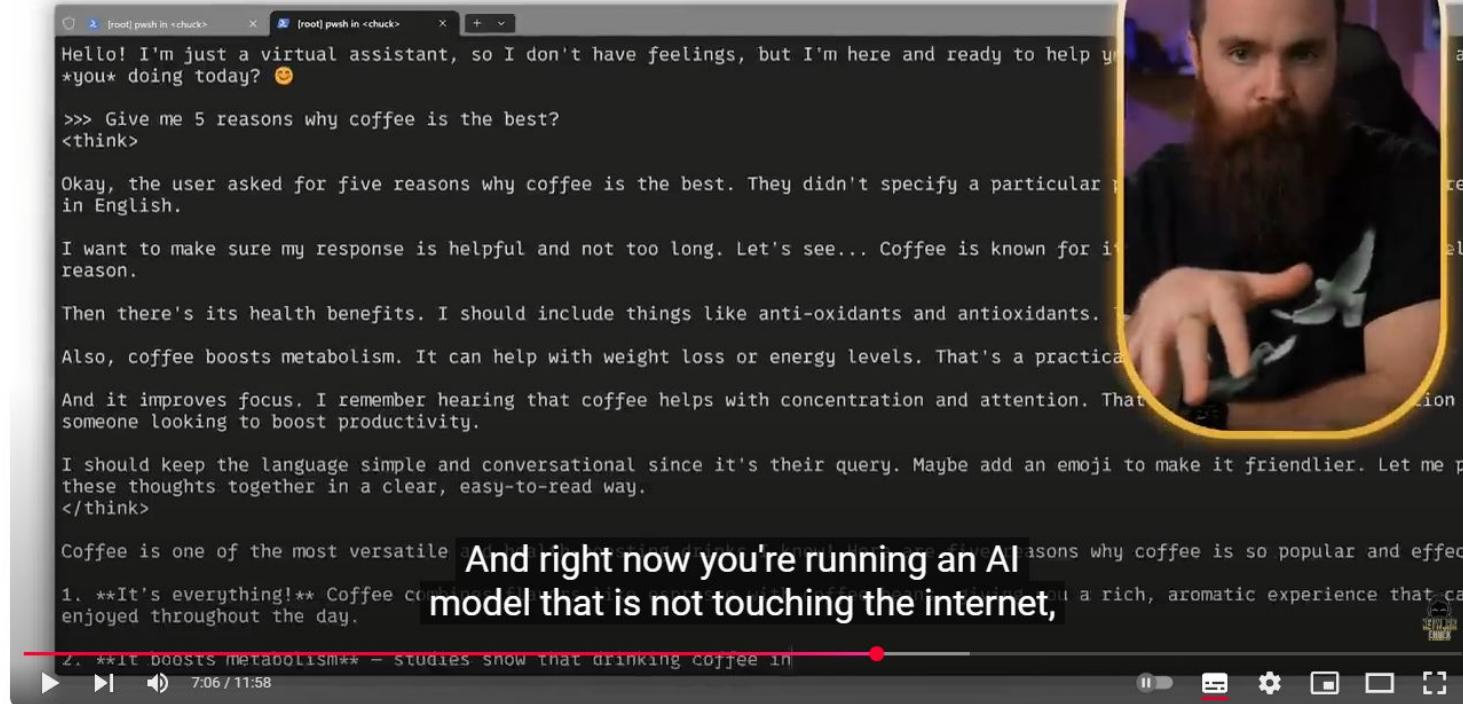


LM Studio interface screenshot showing distillation options:

- Qwen2.5 7B Instruct 1M** • Qwen (7B) - Powerful general purpose instruct model with up to 1 million tokens context.
- DeepSeek R1 Distill (Qwen 7B)** • deepseek (qwen) (7B) - DeepSeek R1 distilled into Qwen 7B: a powerful reasoning model in a small package.
- DeepSeek R1 Distill (Llama 8B)** • deepseek (llama) (8B) - DeepSeek R1 distilled into Llama 8B: a powerful reasoning model in a small package.

Get up and running with large language models.  
Run Llama 3.3, DeepSeek-R1, Phi-4, Mistral, Gemma 2, and other models, locally.

Explore models →  
Available for macOS, Linux, and Windows



Hello! I'm just a virtual assistant, so I don't have feelings, but I'm here and ready to help you doing today? 😊

>>> Give me 5 reasons why coffee is the best?

</think>

Okay, the user asked for five reasons why coffee is the best. They didn't specify a particular reason.

I want to make sure my response is helpful and not too long. Let's see... Coffee is known for its health benefits. Then there's its health benefits. I should include things like anti-oxidants and antioxidants. Also, coffee boosts metabolism. It can help with weight loss or energy levels. That's a practical reason.

And it improves focus. I remember hearing that coffee helps with concentration and attention. That's another reason for someone looking to boost productivity.

I should keep the language simple and conversational since it's their query. Maybe add an emoji to make it friendlier. Let me put these thoughts together in a clear, easy-to-read way.

</think>

Coffee is one of the most versatile and delicious beverages in the world. Here are five reasons why coffee is so popular and effective:

**And right now you're running an AI model that is not touching the internet,**

1. **It's everything!** Coffee combines the perfect blend of flavor and aroma, providing a rich, aromatic experience that can be enjoyed throughout the day.

2. **It boosts metabolism** – studies show that drinking coffee in

▶ ▶ ⏴ 7:06 / 11:58



the ONLY way to run Deepseek...



NetworkChuck  
4,44 mi de inscritos

Seja membro

Inscrito

33 mil



Compartilhar

Download

Valeu

...



4.4  
milhões

<https://www.youtube.com/watch?v=7TR-FLWNVHY>

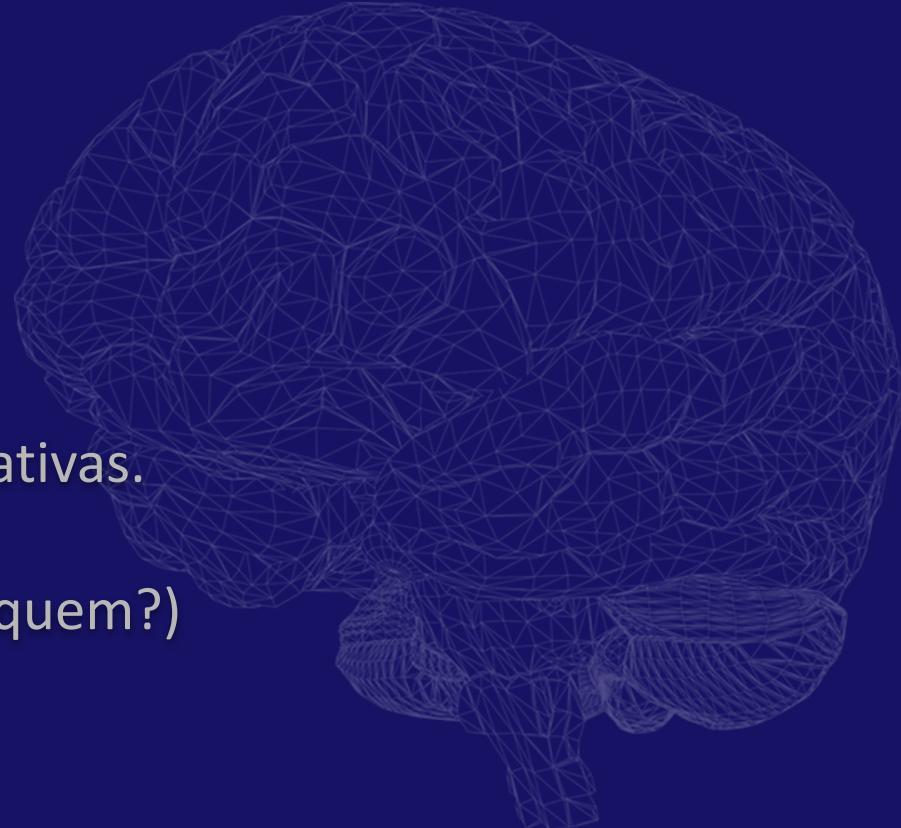
~~Parte 1 - Cases~~

~~Parte 2 - Que tipo de I.A eu preciso?~~

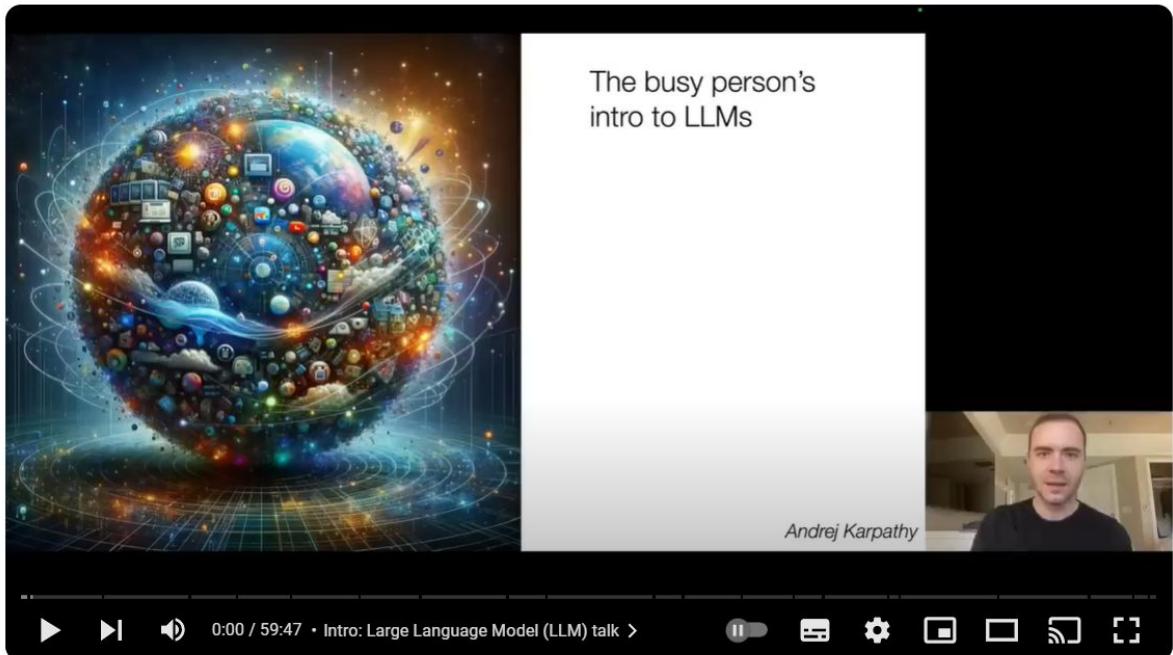
Parte 3 - Avanços Recentes e Expectativas.

BÔNUS 1 - Evolução da I.A (Quem é quem?)

BÔNUS 2 - Onde e o que aprender.



# LLMs



[https://www.youtube.com/  
watch?v=zjkBMFhNj\\_g&t](https://www.youtube.com/watch?v=zjkBMFhNj_g&t)



## [1hr Talk] Intro to Large Language Models



Andrej Karpathy  
427 mil inscritos



Inscrito ▾



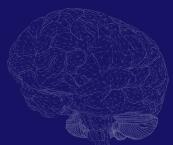
59 mil



0



Compartilhar



# MULTIMODAL LLMs

2-21 Na Fig. 2-20, achar o peso do bloco suspenso sendo a tensão da corda diagonal igual a 20 N.

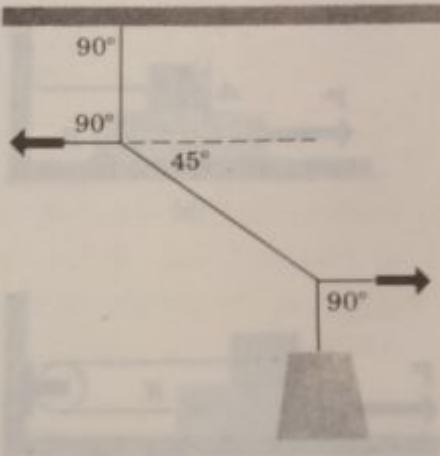


Fig. 2-20

**Resposta:**

O peso do bloco suspenso é aproximadamente  $w = 14.14 \text{ N}$ .



O problema 2-21 pede para calcular o peso do bloco suspenso na Fig. 2-20, onde a tensão na corda diagonal é dada como 20 N, e há um ângulo de 45° entre a corda diagonal e a horizontal.

**Etapas de Resolução:**

**1. Identificar as forças no sistema:**

- A tensão na corda diagonal é  $T = 20 \text{ N}$ .
- O peso do bloco  $w$  atua verticalmente para baixo.
- O sistema está em equilíbrio, então as somas das componentes das forças nas direções horizontal e vertical devem ser zero.

**2. Equilíbrio horizontal:**

As forças horizontais se equilibram. A única força com uma componente horizontal é a tensão da corda diagonal  $T$ . Podemos decompor essa força em sua componente horizontal:

$$T_{\text{horizontal}} = T \cdot \cos(45^\circ)$$

Substituindo  $T = 20 \text{ N}$ :

$$T_{\text{horizontal}} = 20 \cdot \cos(45^\circ) = 20 \cdot \frac{\sqrt{2}}{2} = 14.14 \text{ N}$$

Essa componente horizontal deve ser equilibrada pela força horizontal de compressão (ou reação) na barra horizontal à esquerda, o que significa que não precisamos resolver mais nada nessa direção, pois a barra suporta essa força.

**3. Equilíbrio vertical:**

A componente vertical da tensão diagonal deve equilibrar o peso do bloco. A componente vertical da força  $T$  é:

$$T_{\text{vertical}} = T \cdot \sin(45^\circ)$$

Substituindo  $T = 20 \text{ N}$ :

$$T_{\text{vertical}} = 20 \cdot \sin(45^\circ) = 20 \cdot \frac{\sqrt{2}}{2} = 14.14 \text{ N}$$

Essa componente vertical é responsável por equilibrar o peso  $w$ . Logo:

$$T_{\text{vertical}} = w$$

$$w = 14.14 \text{ N}$$

**Resposta:**

O peso do bloco suspenso é aproximadamente  $\downarrow = 14.14 \text{ N}$ .



ChatGPT 4o ▾

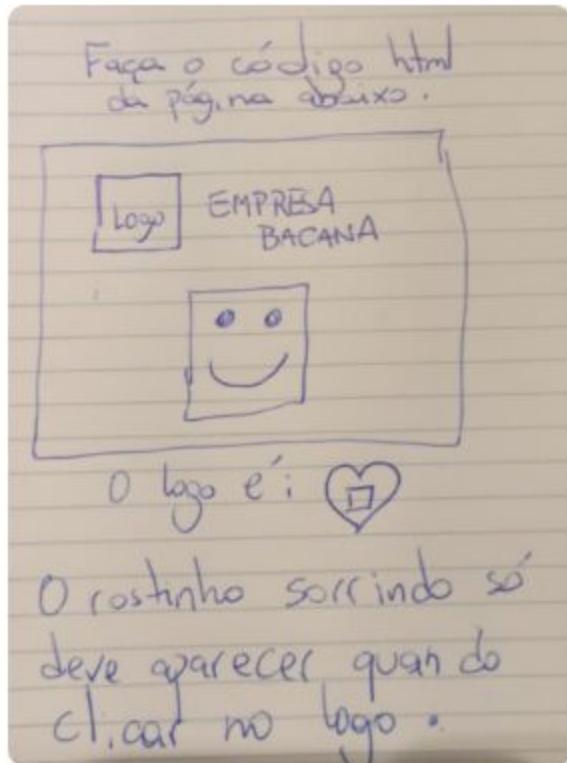


Aqui está um exemplo de código HTML com JavaScript que implementa o que foi solicitado na imagem:

html

Copy code

```
<!DOCTYPE html>
<html lang="pt-BR">
<head>
 <meta charset="UTF-8">
 <meta name="viewport" content="width=device-width,
 <title>Empresa Bacana</title>
 <style>
 body {
 font-family: Arial, sans-serif;
 text-align: center;
 margin-top: 50px;
 }
 #logo {
```



Arquivo

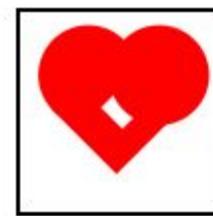
file:///C:/Users/User/...



EMPRESA BACANA

Arquivo

file:///C:/Users/User/...



EMPRESA BACANA



# GERAÇÃO DE CÓDIGO, TESTE E DEBUG



CURSOR

Pricing

Features

Forum

Docs

Careers

Blog

Settings

Logout

Download

## The AI Code Editor

Built to make you extraordinarily productive,  
Cursor is the best way to code with AI.



Download for Free



Watch Demo  
1 Minute

The screenshot shows a Microsoft Visual Studio Code (VS Code) interface. On the left, the sidebar displays a project structure for 'AFRAC\_SEFAPR' containing files like 'flagged', 'MODELS', 'templates', '.gitignore', and numerous CSV files (e.g., '001\_EDA\_(Dados\_dos\_arquivos).csv', '002\_EMBEDDINGS1\_Por\_Itens.csv'). Below the sidebar is the 'OUTLINE' and 'TIMELINE' sections.

The main area features a code editor with a red circle highlighting the 'Code' tab. A yellow circle highlights the 'CHAT' tab at the top right. The code editor contains a Python script named 'app.py' with the following content:

```
all_descs = []
for i, dfg in enumerate(dfs):
 all_descs += [[]]
 try:
 while True:
 df = next(dfg)
 if len(df) == 0:
 break
 else:
 all_descs[-1] = list(set(list(df["DS_ITEM"])))
 print(i+1, len(all_descs[-1]), end="\r")
 except StopIteration:
 None
```

Below the code editor, the status bar shows '[17]', '8m 57.9s', and 'Python'. The bottom navigation bar includes tabs for 'PROBLEMS' (13), 'OUTPUT', and 'TERMINAL', along with a powershell icon.

The 'CHAT' panel on the right contains a message from 'claudie-3.5-sonnet' asking for a Portuguese explanation of the code. The code itself is displayed in the chat window, with a 'Run as cell' button at the bottom right. A detailed analysis of the code follows:

Este código está processando uma série de DataFrames (provavelmente arquivos CSV) para extrair descrições únicas de itens. Vamos analisar em detalhes:

1. `all_descs = []`: Inicializa uma lista vazia para armazenar as descrições.
2. `for i, dfg in enumerate(dfs):`: Itera sobre `dfs`, que parece ser uma lista de geradores de DataFrames (provavelmente criados com `pd.read_csv(..., chunksize=...)`).
3. `all_descs += [[]]`: Adiciona uma nova lista vazia para cada arquivo/gerador.
4. O bloco `try-except` é usado para iterar sobre os chunks de cada

Preciso de um código que

**ANTES**

```
025 - Get only descriptions.ipynb U
Code Markdown Run All Restart ... tf230 (Python 3.7.10)

df = next(dfs)
if len(df) == 0:
 break
else:
 all_descs[-1] = list(set(list(df["DS_ITEM"])))
 print(f"\r{i+1} {len(all_descs[-1])}", end="\r")
except StopIteration:
 None
```

12m 56.6s

**DEPOIS**

```
all_descs = []
for i, dfg in enumerate(dfs):
 unique_descs = set()
 for df in dfg:
 unique_descs.update(df["DS_ITEM"])
 print(f"\r{i+1} {len(unique_descs)}", end="")
 all_descs.append(list(unique_descs))
```

2m 25.0s

1608798

5x mais rápido

CHAT

Este código está efetivamente coletando todas as descrições únicas de itens de vários arquivos CSV, armazenando-as separadamente para cada arquivo na lista `all_descs`. Isso é útil para análise de texto ou para criar um vocabulário único de descrições de produtos.

Você consegue deixar esse código mais curto?

Certamente! Podemos otimizar e encurtar o código da seguinte maneira:

```
all_descs = []
for i, dfg in enumerate(dfs):
 unique_descs = set()
 for df in dfg:
 unique_descs.update(df["DS_ITEM"])
 print(f"\r{i+1} {len(unique_descs)}", end="")
 all_descs.append(list(unique_descs))
```

Run as cell

Explicação das mudanças:

1. Removemos o loop `while True` e o bloco `try-except`, substituindo-os por um simples `for df in dfg`.
2. Usamos `set()` desde o inicio para armazenar descrições únicas, eliminando a necessidade de converter para lista e de volta para set em cada iteração.
3. Utilizamos o método `update()` do set para adicionar novas descrições de forma eficiente.
4. Simplificamos a impressão do progresso usando f-string.

# Dados Sintéticos

Tenho um produto com a descrição "Sabão em pó Omo Multi Ação 400g" gere para mim descrições equivalentes a essa para que eu tenha um banco de sinônimos



Aqui estão algumas descrições equivalentes para "Sabão em pó Omo Multi Ação 400g" que você pode usar como sinônimos:

1. Detergente em pó Omo Multi Ação 400g
2. Sabão em pó Omo Multiuso 400g
3. Sabão em pó Omo Multi Ação 400g embalagem
4. Omo Sabão em pó Multi Ação 400g
5. Detergente em pó Omo para Lavagem Multi Ação 400g
6. Omo Sabão em pó Multiuso 400g
7. Sabão para roupas Omo Multi Ação 400g
8. Sabão Omo em pó Multi Ação 400g
9. Detergente para roupas Omo Multi Ação em pó 400g
10. Omo Multi Ação sabão em pó 400g

+/- 10s pra gerar

Essas variações mantêm o sentido original, enquanto oferecem opções de sinônimos que podem ser usadas em seu banco de dados.

# Dados Sintéticos

jupyter 030 - EDA - NANOGPT\_EMBEDDINGS Last Checkpoint: 2 months ago

File Edit View Run Kernel Settings Help

Code JupyterLab

```
[45]: #context = torch.zeros((1, 1), dtype=torch.long, device=device)
import string

for i in range(30):
 context = torch.tensor([encode(random.choice(string.ascii_uppercase))], dtype=torch.long, device=device)
 print(decode(m.generate(context, max_new_tokens=200)[0].tolist()).split("\n")[0])
```

Creme Alis Aorgan300  
XIC CAFE FLORENCA 4PECAS  
MESA REF107/360  
WAFFER BARION FLOCOS/AVELA PCT 35G  
COMFORTIS CAES ADULT -  
Leite Monange] Semi-  
LIDOCAINA 2% COLIRIO DUOM  
FAR.MAND.YOK'Semidesn.500g  
X REPARO SC TRA NXR 125/150 - PARTES E PECAS 76 METAL  
KT3 CALC 2831798 DUBAI  
KIT APAR BARB ACQUA  
SHOYU CEPERA1LT



# LMSYS Chatbot Arena (Multimodal)

Eu não tenho  
o Chatgpt  
pago?

<https://chat.lmsys.org/>



Total #models: 201. Total #votes: 2,634,992. Last updated: 2025-02-09.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [larena.ai](https://larena.ai)!

Category: Overall | Apply filter:  Style Control |  Show Deprecated

Overall Questions: #models: 201 (100%) #votes: 2,634,992 (100%)

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	+5/-6	11462	Google	Proprietary
1	1	Gemini-2.0-Pro-Exp-02-05	1379	+5/-8	9385	Google	Proprietary
3	1	ChatGPT-4a-latest-(2024-11-20)	1365	+3/-3	39649	OpenAI	Proprietary
3	1	DeepSeek-R1	1361	+9/-9	4193	DeepSeek	MIT
4	7	Gemini-2.0-Flash-001	1355	+5/-8	7264	Google	Proprietary
4	1	o1-2024-12-17	1351	+5/-6	13416	OpenAI	Proprietary
7	6	o1-preview	1335	+4/-4	33164	OpenAI	Proprietary
7	7	Qwen2.5-Max	1332	+7/-9	5459	Alibaba	Proprietary



# LMSYS Chatbot Arena (Multimodal)

👉 Chat now!

🔍 Expand to see the descriptions of 73 models

Model A



Pra quem o cachorro está olhando?

O cachorro está olhando para a pessoa sentada no sofá à direita.

Model B



Pra quem o cachorro está olhando?

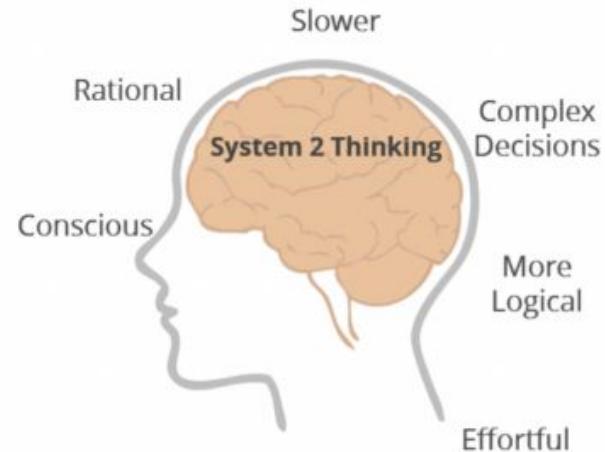
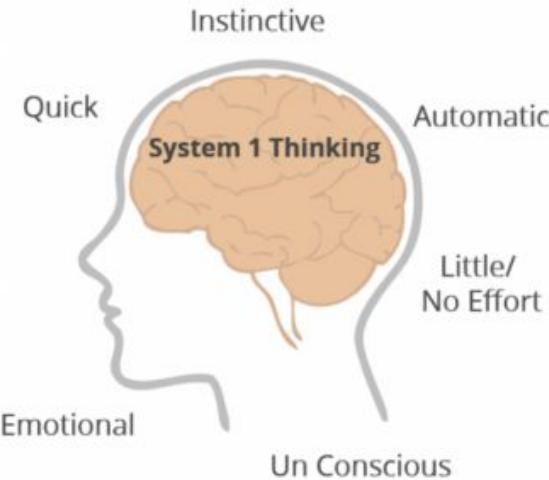
Na imagem, o cachorro está olhando para a mulher que está sentada no sofá.

👉 A is better

👉 B is better

之心 Tie

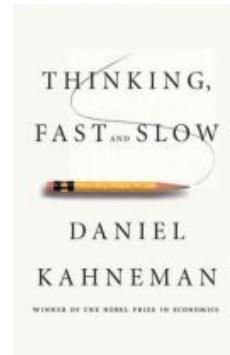
👎 Both are bad



$$2 + 2 =$$

$$17 \times 24 =$$

A maior parte dos modelos de I.A.  
(incluindo LLMs) são análogas ao  
sistema 1.



Fale sobre o evento da BX Brasil

Fale sobre o evento da BX Brasil 2025

Found 37 results ▶

Thought for 41 seconds ▲

Okay, o usuário quer saber sobre o evento BX Brasil 2025. Vou começar verificando os resultados de busca que tenho. Vejo que as páginas 1, 2 e 4 são diretamente relacionadas ao evento BX Brasil. As outras páginas tratam de outros temas, como mercado financeiro, eventos de marketing e o portal BRICS. Então, preciso focar nas informações das páginas 1, 2 e 4.

A página 1 detalha o evento, mencionando que é a 3ª edição, datas (13 e 14 de fevereiro de 2025), local provavelmente não especificado, mas fala em indicação de hotéis próximos. Também tem informações sobre o conteúdo do evento, como palestras sobre IA e estratégias empresariais, networking, e inclusões como material de apoio e coffee break. Além disso, menciona o cupom de desconto "EVNTOS".

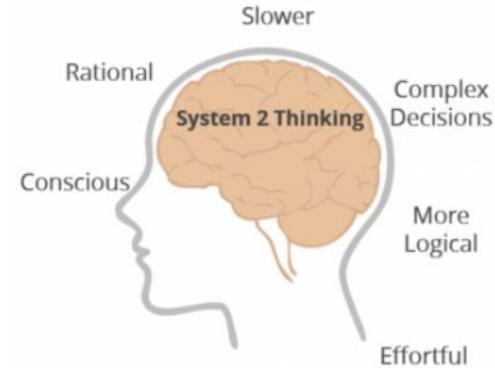
A página 2 parece ser a homepage do evento, com programação detalhada dos dois dias. Dia 1 tem palestras sobre vendas, IA, investimentos, e um tour na TOTVS. Dia 2 aborda lucro máximo, gestão de times com IA, tendências tecnológicas e painéis. Também confirma os palestrantes e detalha o que está incluso no ingresso, como acesso a dois dias, material, coffee break e água. Além disso, menciona que as apresentações serão disponibilizadas e que o evento ocorre em locais diferentes cada dia.

A página 4 é a página de ingressos, reforçando a data, informações de contato, e detalhes sobre inscrição, como parcelamento em até 3x sem juros. Também menciona que participantes de todos os níveis podem participar e que a organização oferece certificado.

Message DeepSeek

DeepThink (R1)

Search



$$17 \times 24 =$$

Finalmente, garantir que a resposta esteja em português, bem estruturada, com citações adequadas e sem informações desnecessárias de outras páginas não relacionadas.

O **BX Brasil 2025** é um evento de imersão executiva focado em **inteligência artificial (IA)** e suas aplicações no ambiente empresarial, promovendo conexões estratégicas e conteúdo de alto nível para profissionais e líderes do mercado. Seguem os principais detalhes:

## 1. Data e Local

- Datas:** 13 e 14 de fevereiro de 2025 (quinta e sexta-feira) 1 3.
- Horário:** Das 8h às 21h, com programação extensa ao longo do dia 3.
- Local:** O evento ocorrerá em diferentes locais a cada dia, conforme a programação divulgada. A organização oferece indicações de hotéis próximos com descontos exclusivos usando o cupom "**EVNTOS**" 1 3.

# Agentic A.I.



Andrew Ng On AI Agentic Workflows And Their Potential For Driving AI Progress

DATA CLOUD SUMMIT 24



## ChatDev: Communicative Agents for Software Development

Chen Qian\* Wei Liu\* Hongzhang Liu\* Nuo Chen\* Yufan Dang\*  
Jiahao Li\* Cheng Yang\* Weize Chen\* Yusheng Su\* Xin Cong\*  
Juyuan Xu\* Dahai Li\* Zhiyuan Liu\*\* Maosong Sun\*\*

\*Tsinghua University \*The University of Sydney \*\*BUPT \*Modelbest Inc.  
qianc62@gmail.com liuzy@tsinghua.edu.cn sms@tsinghua.edu.cn

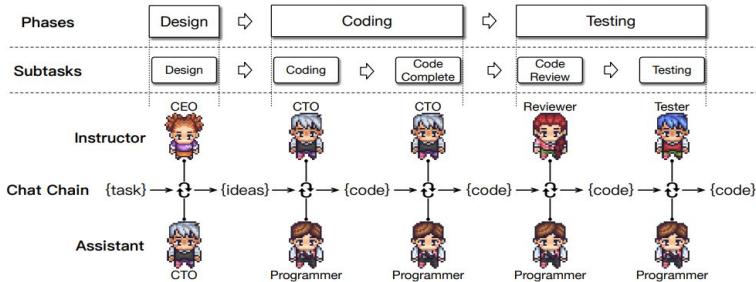
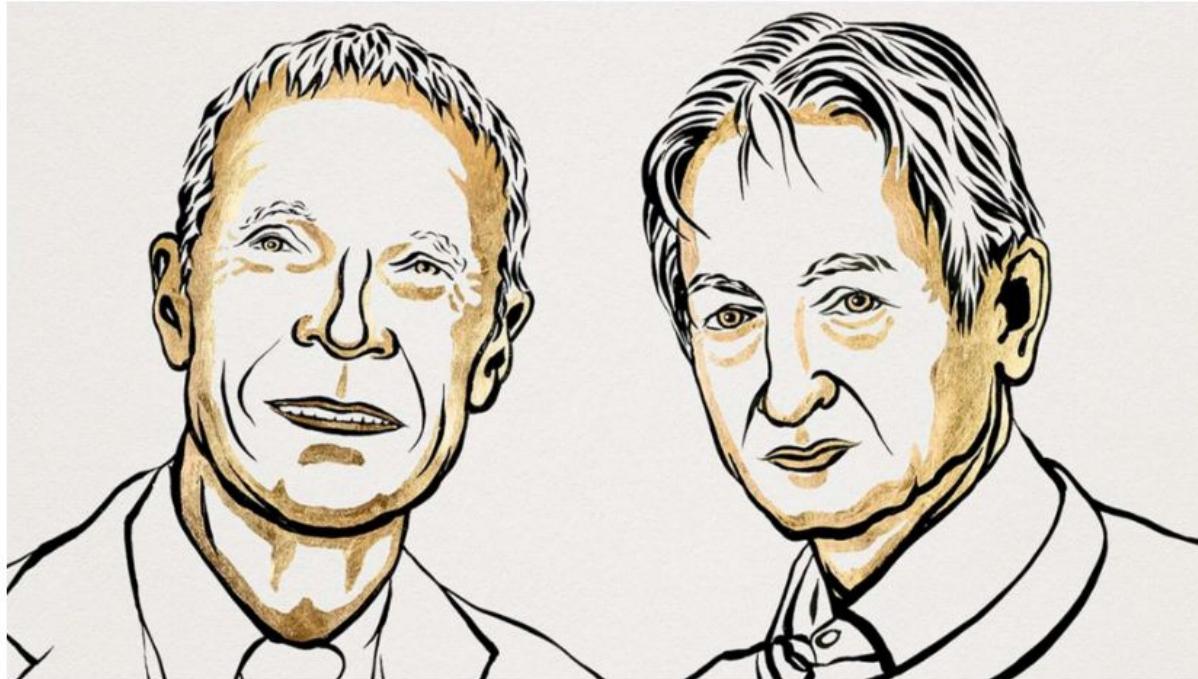


Figure 1: ChatDev, a chat-powered software development framework, integrates LLM agents with various social roles, working autonomously to develop comprehensive solutions via multi-agent collaboration.

# Pioneiros da inteligência artificial ganham Nobel de Física 2024

A dupla John Hopfield e Geoffrey Hinton ganhou o Nobel de Física de 2024 pelos seus trabalhos voltados ao aprendizado de máquina; entenda

Pedro Spadoni | 08/10/2024 08h44



John Hopfield (esq.) e Geoffrey E. Hinton (dir.) (Imagem: Niklas Elmehed/Nobel Assembly)

# Nobel de Química 2024 vai para trio que usou IA para decifrar proteínas

David Baker, Demis Hassabis e John Jumper ganharam Nobel de Química 2024 por decifrarem segredos das proteínas por meio da computação e IA

Pedro Spadoni | 09/10/2024 08h46



David Baker (esq.), Demis Hassabis (meio) e John Jumper (dir.) (Imagem: Niklas Elmehed/Nobel Assembly)

## How Long Until Computers Have the Same Power As the Human Brain?

Lake Michigan's volume (in fluid ounces) is about the same as our brain's capacity (in calculations per second). Computing power doubles every 18 months. At that rate, you see very little progress for a long time—and suddenly you're finished.



<https://www.motherjones.com/media/2013/05/robots-artificial-intelligence-jobs-automation/>



***“Just as 100 years ago electricity transformed industry after industry, AI will now do the same.” – Andrew Ng***

*Andrew Ng is Founder & CEO of Landing AI, Founder of deeplearning.ai, Co-Chairman and Co-Founder of Coursera, and is currently an Adjunct Professor at Stanford University. He was also Chief Scientist at Baidu Inc., and Founder & Lead for the Google Brain Project*



**Marcelo Filipak**

filipak@filiart.com.br

<https://www.linkedin.com/in/marcelo-filipak-97386710/>

<https://github.com/mfilipak>

# I.A. NA PRÁTICA

cases em automação comercial

**OBRIGADO!**



# I.A. NA PRÁTICA

cases em automação comercial

BAIXE O MATERIAL

OBRIGADO!

**Marcelo Filipak**

filipak@filiart.com.br

<https://www.linkedin.com/in/marcelo-filipak-97386710/>

<https://github.com/mfilipak>

# Um lugar para aprender ou se atualizar. (MIT Courses)

<http://introtodeeplearning.com/>



## Schedule

New 2024 lectures, slides, and labs!  
*Taught in-person at MIT — open-sourced to the world.*

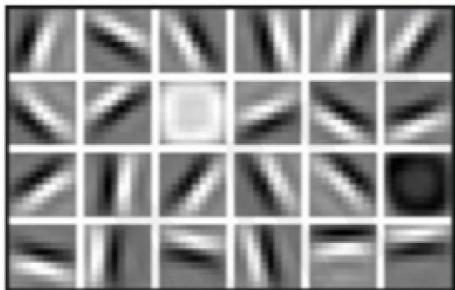
<b>Intro to Deep Learning</b> Lecture 1 Apr. 29, 2024 [Slides] [Video]	<b>Deep Sequence Modeling</b> Lecture 2 May 6, 2024 [Slides] [Video]	<b>Intro to TensorFlow; Music Generation</b> Software Lab 1 [Code]
<b>Deep Computer Vision</b> Lecture 3 May 13, 2024 [Slides] [Video]	<b>Deep Generative Modeling</b> Lecture 4 May 20, 2024 [Slides] [Video]	<b>Facial Detection Systems</b> Software Lab 2 [Paper] [Code]
<b>Deep Reinforcement Learning</b> Lecture 5 May 27, 2024 [Slides] [Video]	<b>New Frontiers</b> Lecture 6 June 3, 2024 [Slides] [Video]	<b>Large Language Models</b> Software Lab 3 [Code] coming soon!
<b>Generative AI for Media</b> Lecture 7 June 10, 2024 [Slides] [Video]	<b>Stories from Models in the Wild</b> Lecture 8 June 17, 2024 [Slides] [Video]	<b>Final Project</b> Work on final projects
<b>Final Project</b> Work on final projects June 24, 2024	<b>Project Presentations</b> Pitch your ideas! June 24, 2024	<b>Awards Ceremony</b> Final awards and celebration!

# Why Deep Learning?

Hand engineered features are time consuming, brittle, and not scalable in practice

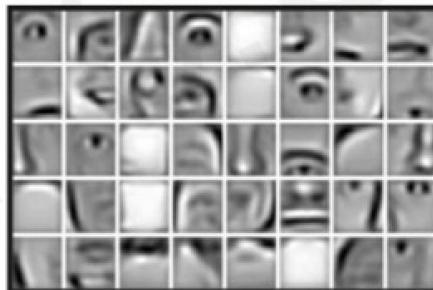
Can we learn the **underlying features** directly from data?

Low Level Features



Lines & Edges

Mid Level Features



Eyes & Nose & Ears

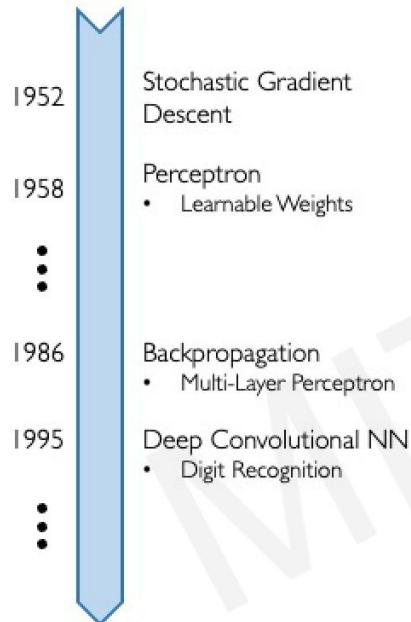
High Level Features



Facial Structure



# Why Now?



Neural Networks date back decades, so why the dominance?

## I. Big Data

- Larger Datasets
- Easier Collection & Storage



## 2. Hardware

- Graphics Processing Units (GPUs)
- Massively Parallelizable



## 3. Software

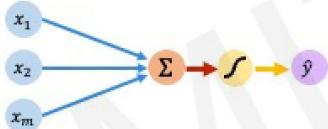
- Improved Techniques
- New Models
- Toolboxes



# Core Foundation Review

## The Perceptron

- Structural building blocks
- Nonlinear activation functions



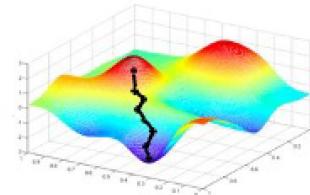
## Neural Networks

- Stacking Perceptrons to form neural networks
- Optimization through backpropagation

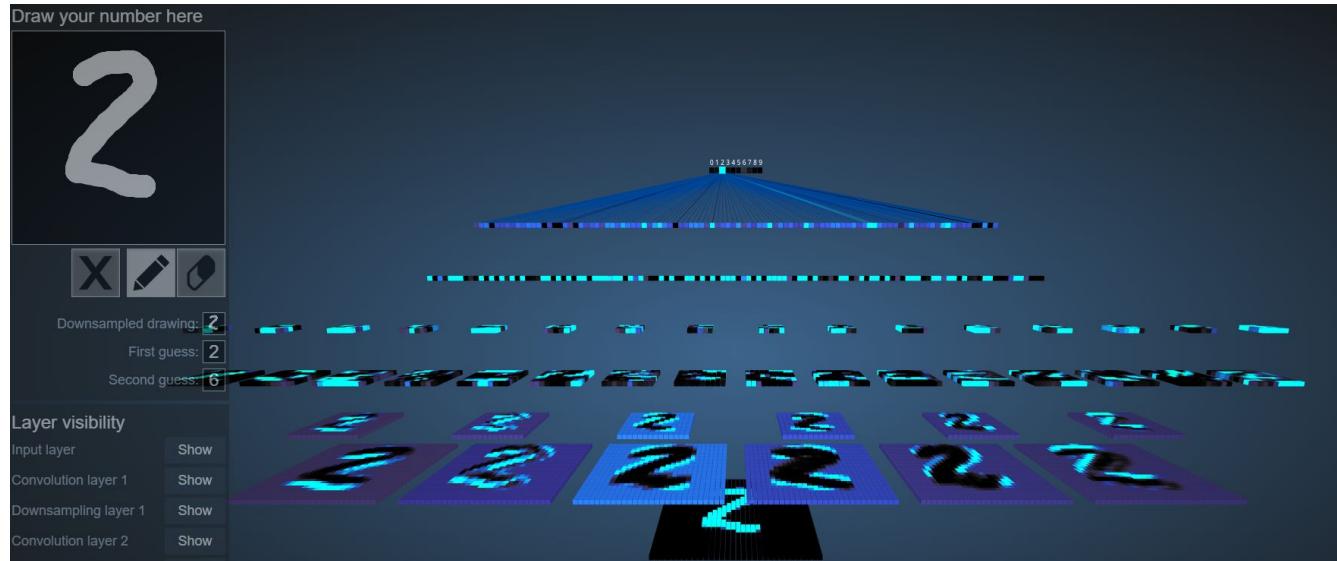


## Training in Practice

- Adaptive learning
- Batching
- Regularization



# Visualização de uma rede neural



[https://adamharley.com/nn\\_vis/cnn/3d.html](https://adamharley.com/nn_vis/cnn/3d.html)

