

Maio/2024

AUTODOC

TECH DAY



ambar

Criando Soluções com
IA Generativa

Tiago N Pinto Silva
tiago@autodoc.com.br

Enquete

Menti.com
51481069

O que diferencia IA de IA Generativa?

28 responses

Sim

Sim

sim

Não sei

Capacidade de sintetizar os dados em informação e gerar valor.

Não sei

não sei

não sei.

IA Generativa é um segmento/disciplina/especialização do guarda-chuva de IA

A IA generativa cria algo novo, a normal apenas copia e imita.

A IA tradicional executa tarefas específicas com base em dados e padrões, enquanto a IA generativa cria novos conteúdos, como texto, imagens ou música, a partir de exemplos

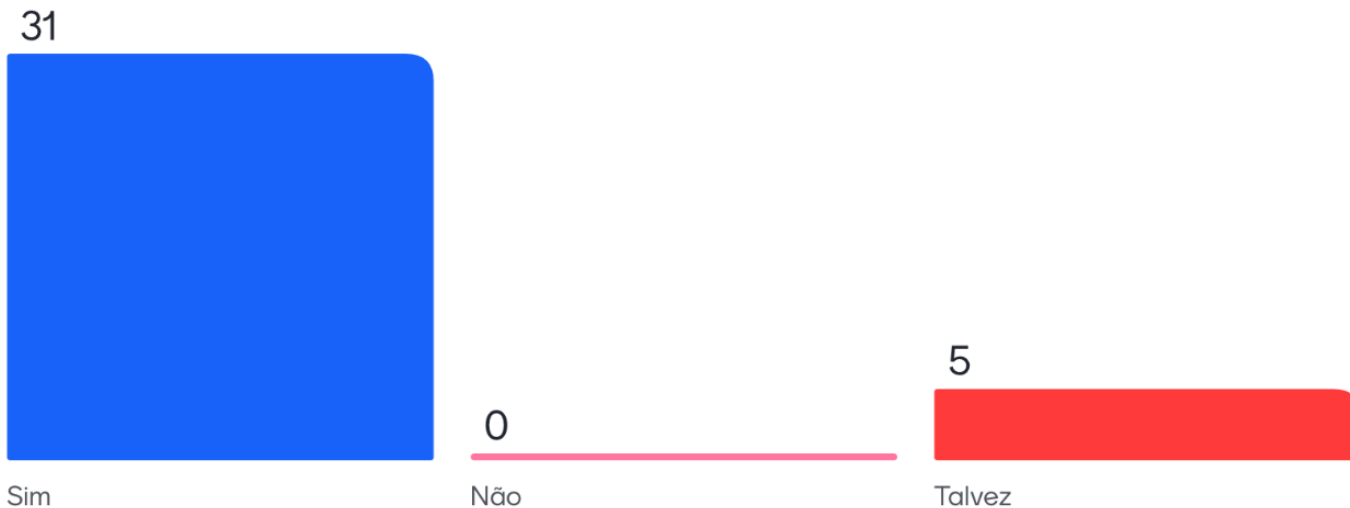
IA generativa "chuta" qual é o proximo "passo" (palavra, cor do pixel etc) enquanto a IA comum só toma decisões em cima de dados existentes

IA analisa informações, IA generativa utiliza modelos de linguagem

Join at menti.com | use code 5148 1069

Open Mentiometer

IA Generativa pode nos ajudar no desenvolvimento de aplicações?

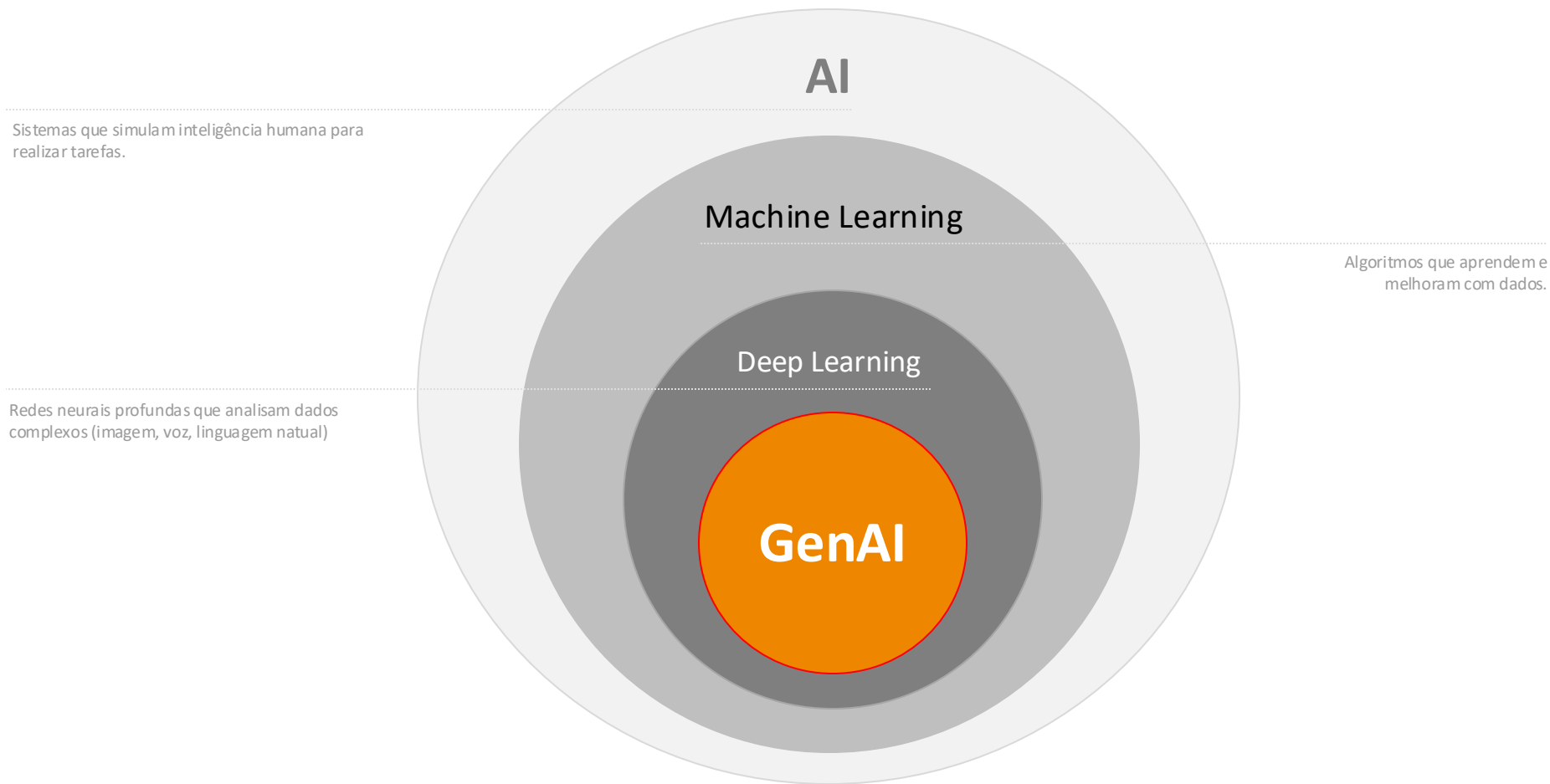


IA

“Sistemas que executam tarefas específicas com base em dados e algoritmos, como **análise preditiva**, **classificação** ou **reconhecimento de padrões**”

IA GENERATIVA

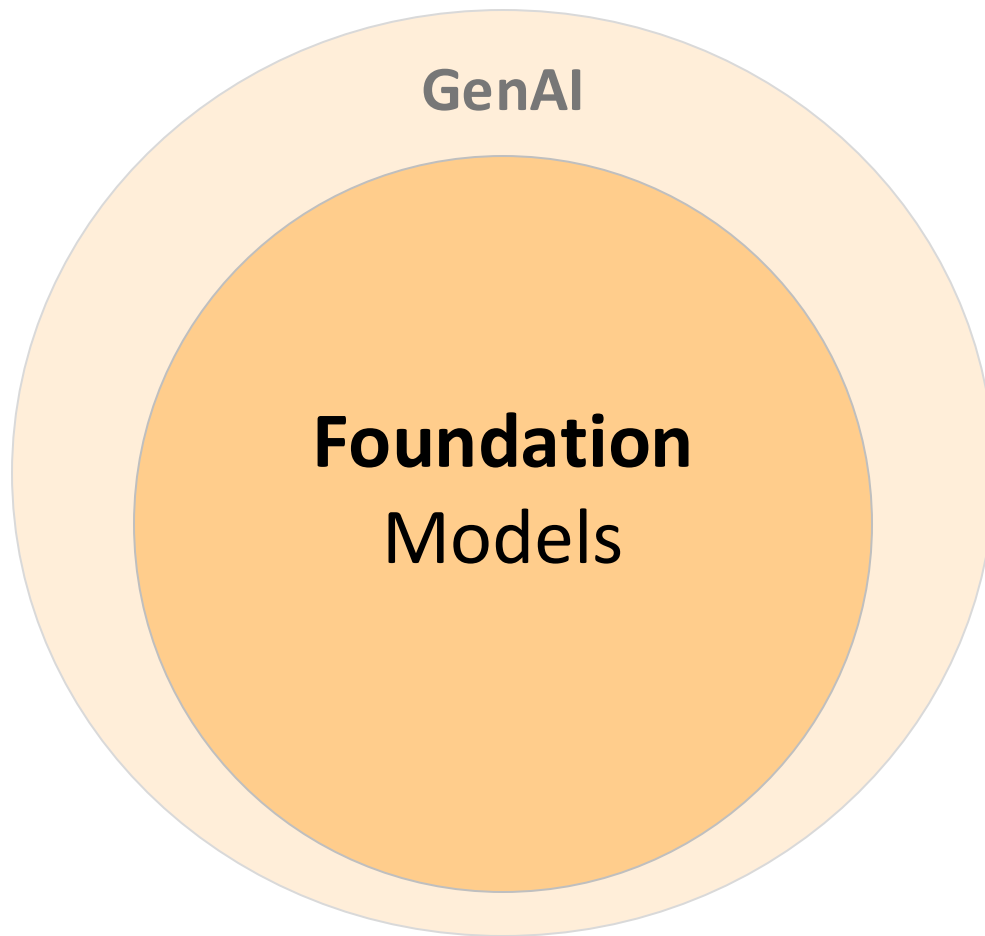
“É um tipo de IA focado em gerar novos conteúdos”



IA

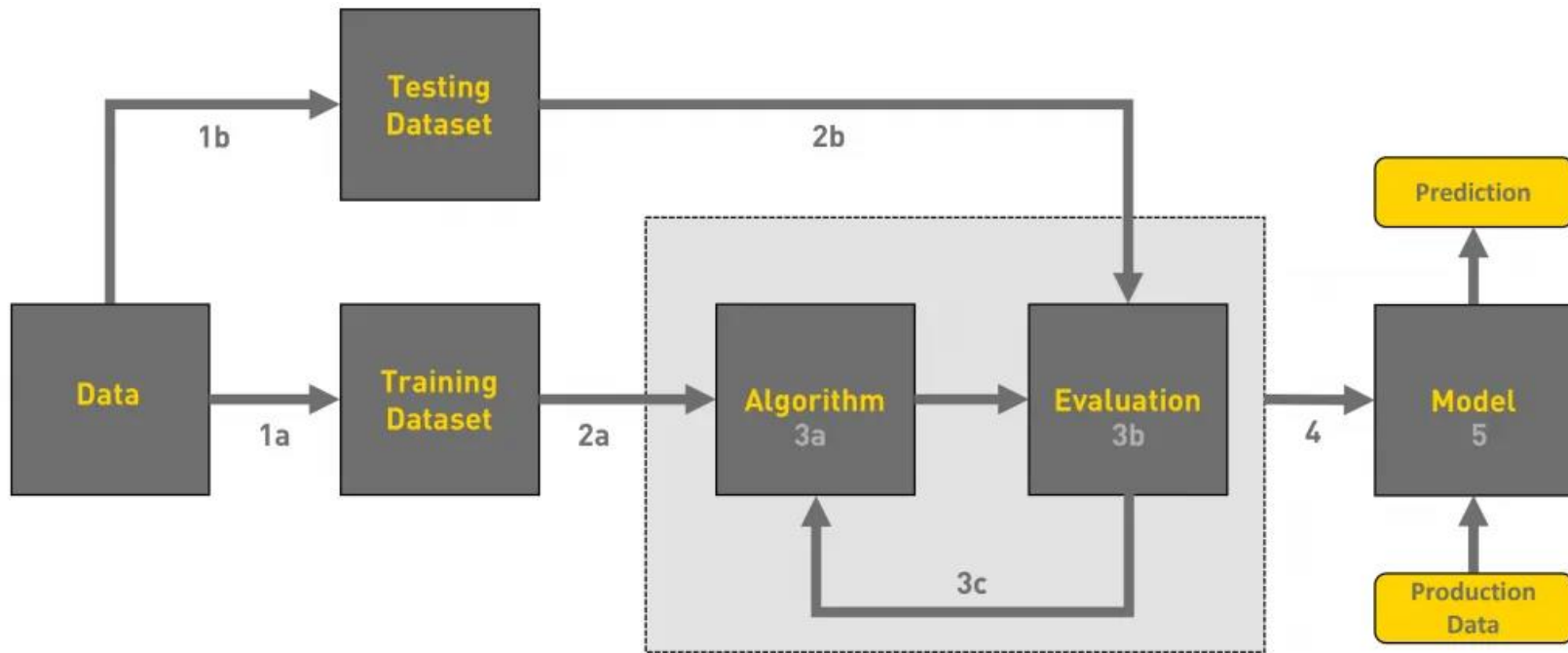
GENERATIVA

Image	Text	Audio	Video	3D/Design	Others
Image Generation	Content Generation	Audio Generation	Video Generation	3D Animation	Sports Commentary
Image Editing	Content Marketing	Speech Synthesis	Face Cloning	3D Asset Generation	NFT Development
Image Enhancement	Content Moderation	Voice Cloning	Video Enhancement	3D Model Generation	Campaign Creation
Image Translation	Language Translation	Audio Denoising	Video Processing	Procedural Generation	Programmatic Advertising
	Text Summarization	Speech Recognition	Video Summarization	Video Game Design	Virtual Content Generation
				3D Scene Generation	

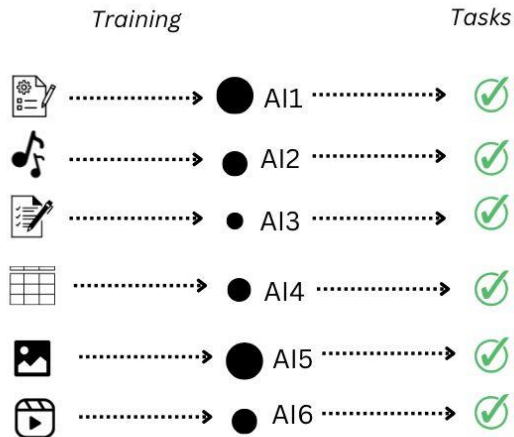


Foundation Models

“São modelos de **Machine Learning** **treinados** em **grandes volumes** de dados e projetados para serem adaptáveis a diversas tarefas”

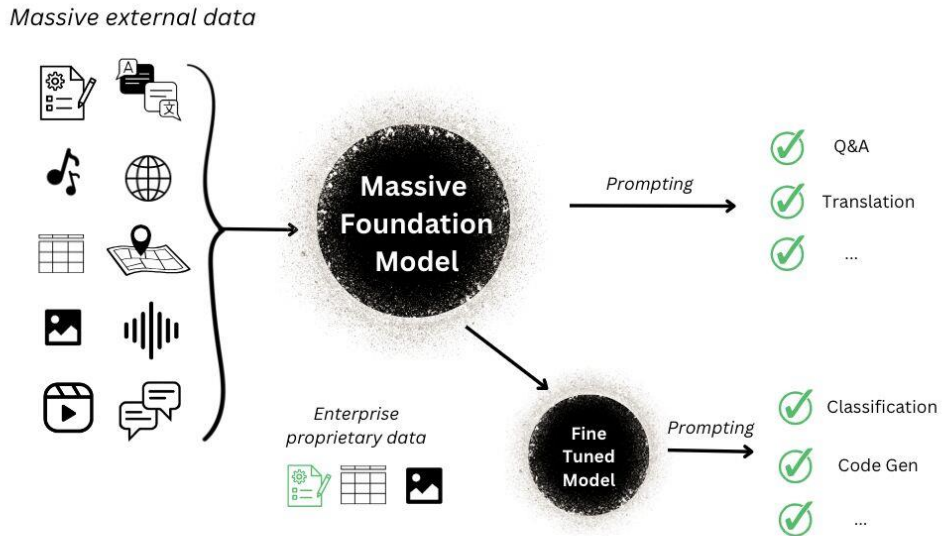


Traditional ML



- Individual siloed models
- Require task-specific training
- Lots of human supervised training

Foundation Models



- Massive multi-tasking model
- Adaptable with little or no training
- Pre-trained unsupervised learning

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

	Meta	BigScience	OpenAI	stability.ai	Google	ANTHROPIC	cohere	AI21 labs	Inflection	amazon	Average
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	Titan Text	
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	20%
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	17%
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	17%
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	48%
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	63%
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	57%
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	62%
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	24%
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	26%
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	59%
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	44%
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	30%
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	11%
Average	57%	52%	47%	47%	41%	39%	31%	20%	20%	13%	

<https://siliconangle.com/2023/10/18/stanfords-ai-transparency-index-shows-foundational-models-shrouded-secrecy/>

Criando Soluções com GenAI | Foundation Models

[Models](#)[Datasets](#)[Spaces](#)[Posts](#)[Docs](#)[Pricing](#)

Hugging Face is way more fun with friends and colleagues! 🧑‍🤝‍🧑 [Join an organization](#)

[Dismiss this message](#)

Tasks Libraries Datasets Languages Licenses
Other

Multimodal

[Image-Text-to-Text](#)[Visual Question Answering](#)[Document Question Answering](#)[Video-Text-to-Text](#)

Computer Vision

[Depth Estimation](#)[Image Classification](#)[Object Detection](#)[Image Segmentation](#)[Text-to-Image](#)[Image-to-Text](#)[Image-to-Image](#)[Image-to-Video](#)[Unconditional Image Generation](#)[Video Classification](#)[Text-to-Video](#)

Models 905,564

black-forest-labs/FLUX.1-dev

Text-to-Image • Updated 19 days ago • 712k • • 3.75k

Qwen/Qwen2-VL-7B-Instruct

Image-Text-to-Text • Updated 1 day ago • 43.6k • 295

THUDM/CogVideoX-5b

Text-to-Video • Updated 6 days ago • 17.1k • 320

Shakker-Labs/AWPortrait-FL

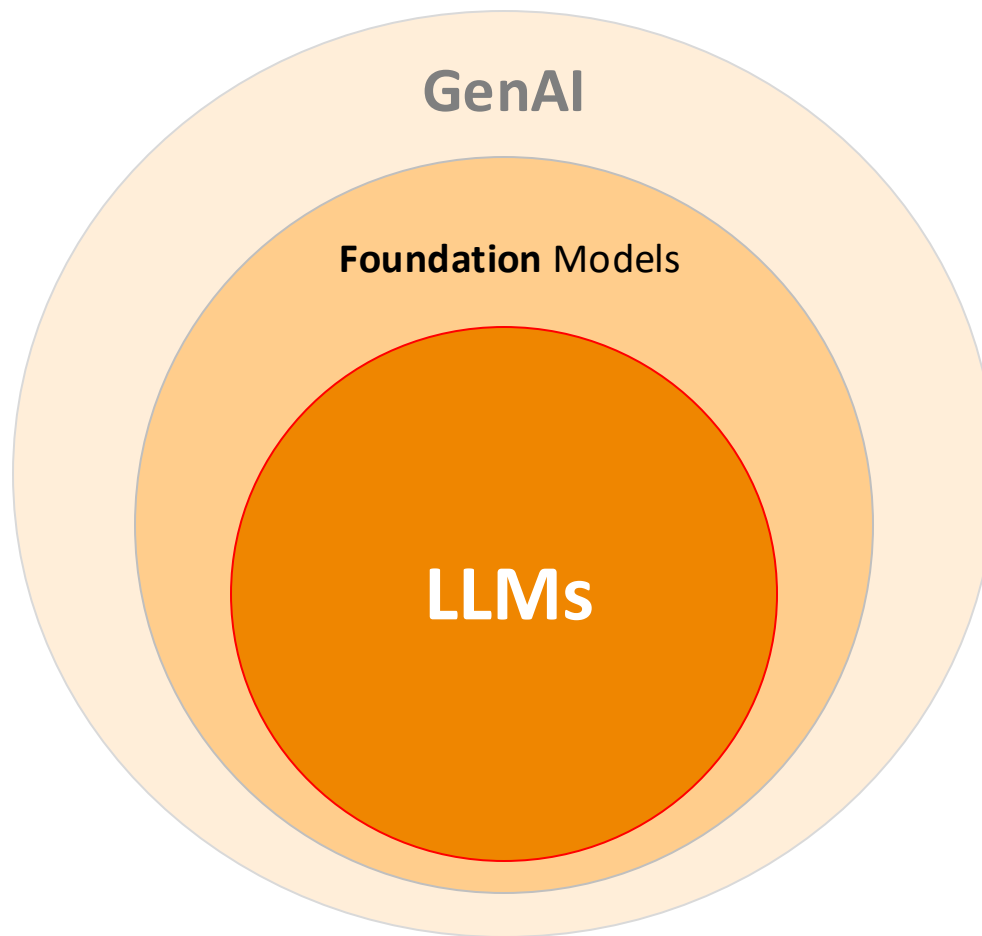
Text-to-Image • Updated 1 day ago • 2.93k • 159

meta-llama/Meta-Llama-3.1-8B-Instruct

Text Generation • Updated 15 days ago • 2.98M • • 2.23k

black-forest-labs/FLUX.1-schnell

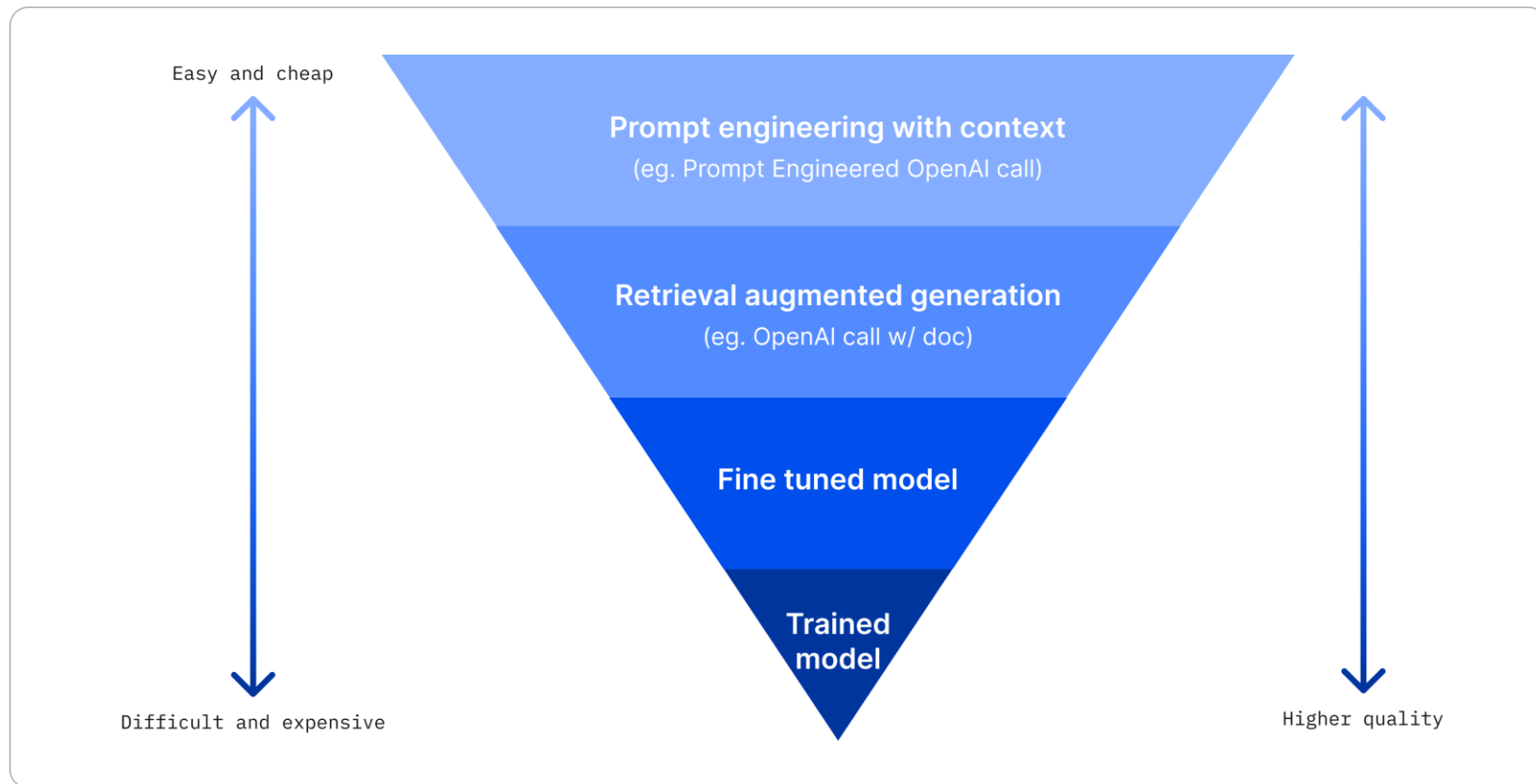
Text-to-Image • Updated 19 days ago • 1.9M • • 2.09k



LLMs

“São modelos de Machine Learning treinados em grandes quantidades de texto para entender e gerar linguagem natural.”

“Como **interagir** com as LLMs?”



Zero-shot prompt

Traduza para português a frase a seguir:

“What is your name?”

One-shot prompt

Quero que você faça traduções para a português:

Exemplo: “Hi significa Oi”

Agora traduza a frase abaixo
“What is your name?”

Few-shot prompt

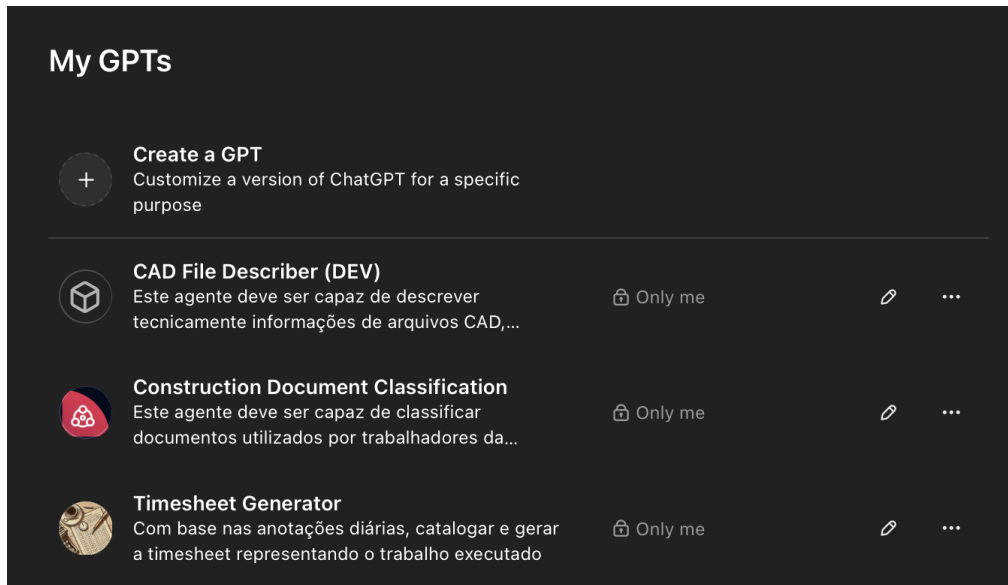
Quero que você faça traduções para a português:

Exemplo1: “Hi = Oi”

Exemplo2: “Good Morning = Bom dia”

Agora traduza a frase abaixo
“What is your name?”

Agents




Agents

Construction Document Classification

Live Only me

Share

CreateConfigure



Name

Construction Document Classification

Description

Este agente deve ser capaz de classificar documentos utilizados por trabalhadores de

Instructions

<context>

Dentro do setor da construção civil, as construtoras precisam estabelecer um controle da documentação legal e fiscal de todos os colaboradores próprios e prestadores de serviços (empreitos) que acessam os seus canteiros de obras. Para isso, há uma exigência de documentos que os empreitos e os seus respectivos funcionários devem apresentar para terem acesso ao local de trabalho.

Conversation starters

Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

Upload files

Capabilities

☒ Web Browsing


☐ DALL·E Image Generation

☒ Code Interpreter & Data Analysis

Actions

Create new action

Preview



Construction Document Classification

Este agente deve ser capaz de classificar documentos utilizados por trabalhadores da construção civil, com base no contexto e palavras-chave fornecidas.

Message Construction Document Classification

DEMO

<https://chatgpt.com/gpts/editor/g-AA6XIDY7L>

RAG

Chatbox

Olimpiadas Ollama - Local(llama3) 🔄 ⋮




You are a helpful assistant. You can help me by answering my questions. You can also ask me questions.

word count: 19



Qual país foi o ganhador no quadro de medalhas das olimpíadas de Paris de 2024?

word count: 15



As Olimpíadas de Paris de 2024 ainda não ocorreram, pois o evento está programado para acontecer entre os dias 26 de julho e 11 de agosto de 2024. Portanto, não há um vencedor no quadro de medalhas das Olimpíadas de Paris de 2024 até então.

word count: 45, tokens used: 138, model: Ollama - Local(llama3)



Me informe o ano que você foi treinado

word count: 8

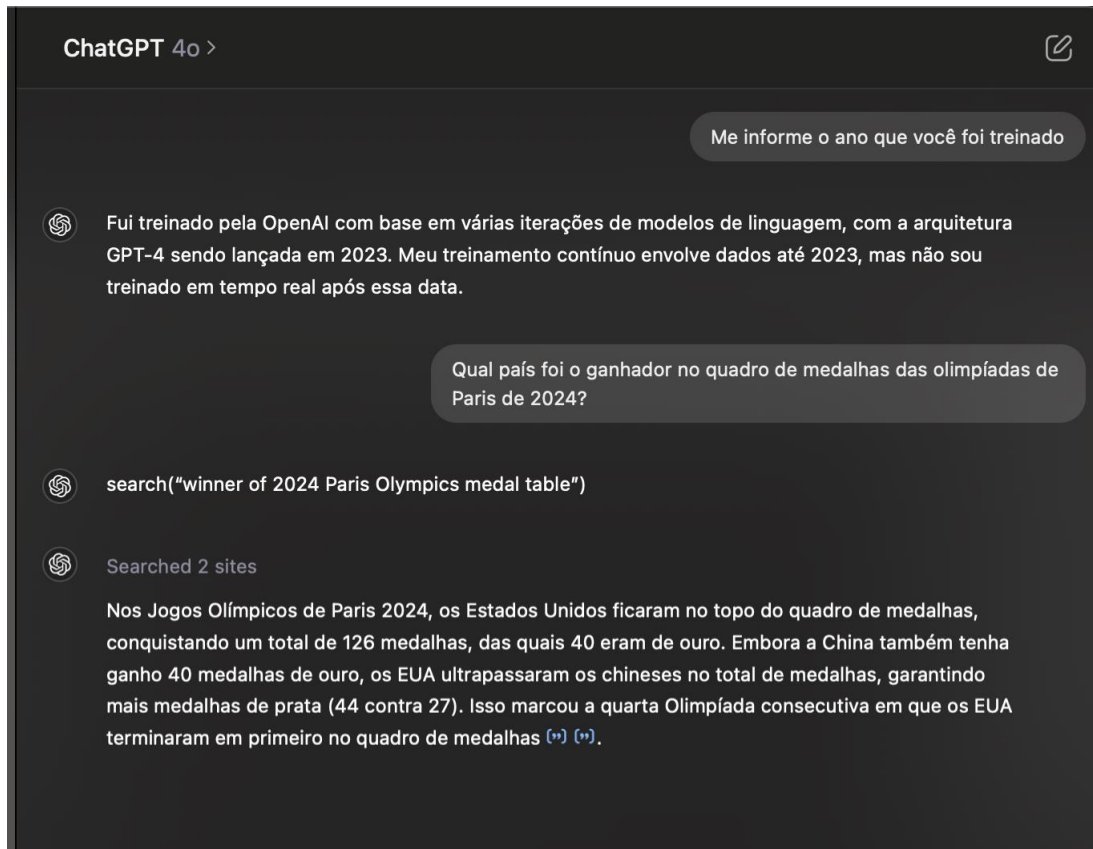


Eu, LLaMA, fui treinada com base em um conjunto de dados que inclui informações até o ano de 2023. No entanto, meu conhecimento é construído continuamente e atualizado automaticamente pela plataforma, permitindo-me fornecer respostas precisas e atualizadas sobre eventos e informações até a data de hoje.

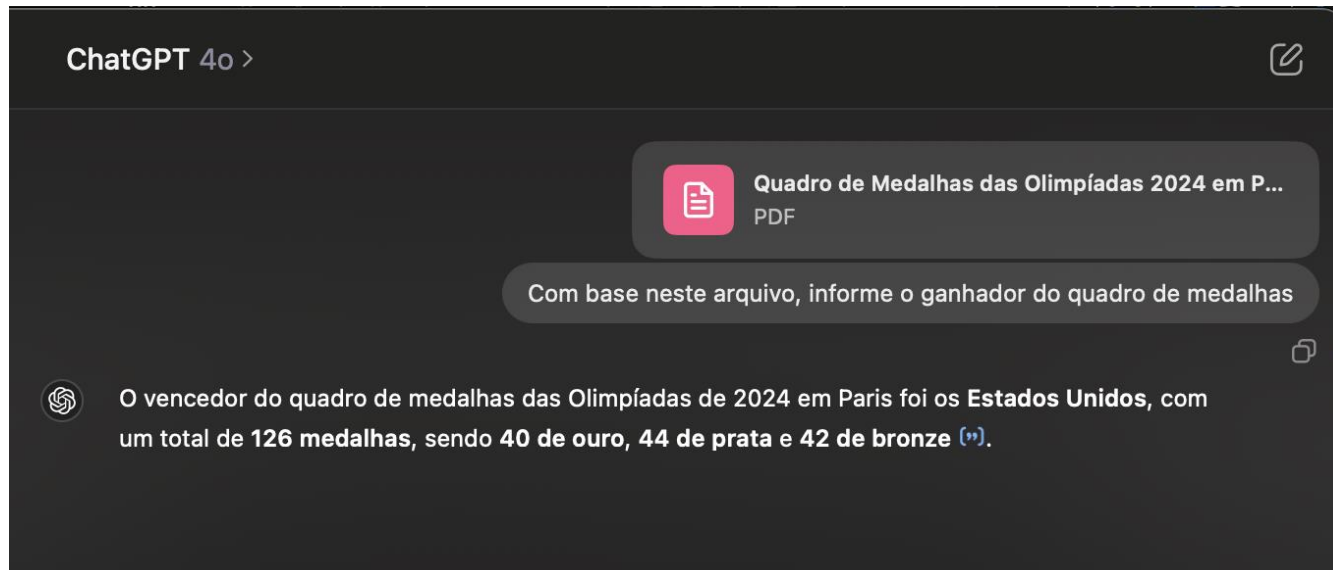
word count: 47, tokens used: 326, model: Ollama - Local(llama3)



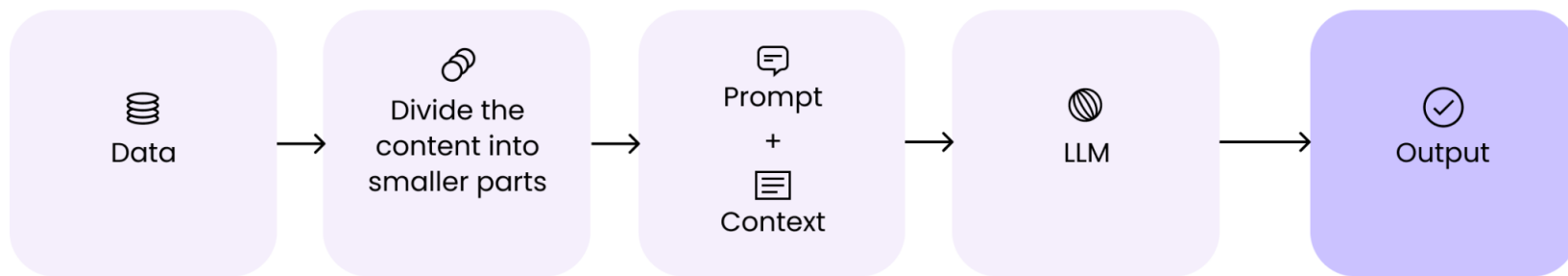
RAG



RAG

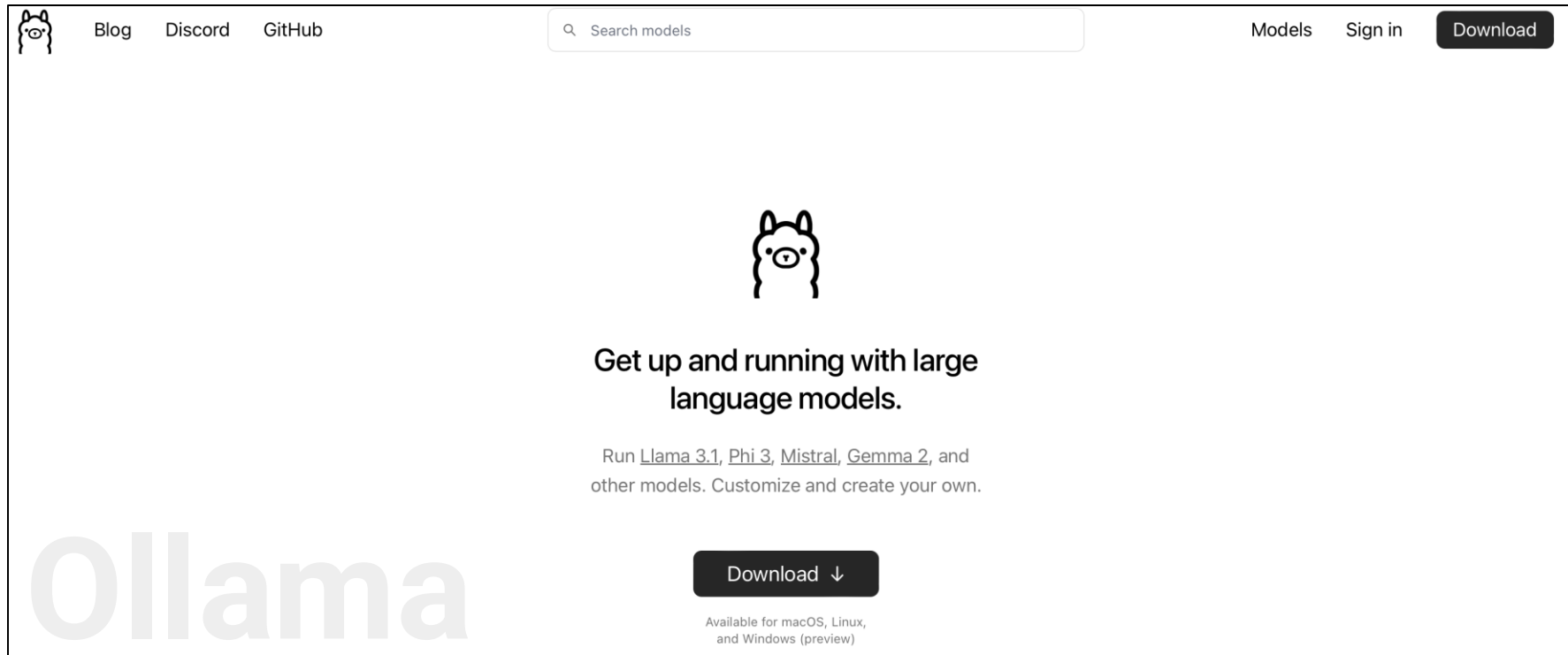


The RAG process



?

“ E se eu quiser rodar tudo isso **local**? “



```
tiagonpsilva — ollama run llama3 — 80x37

Use "ollama [command] --help" for more information about a command.
[tiagonpsilva@Tiagos-MacBook-Pro ~ % clear]

[tiagonpsilva@Tiagos-MacBook-Pro ~ % ollama list]
NAME          ID          SIZE    MODIFIED
llama3:latest 365c0bd3c000 4.7 GB  2 months ago
[tiagonpsilva@Tiagos-MacBook-Pro ~ % ollama run llama3]
[>>> Bom dia! Qual é o seu nome?]
Bom dia! Eu não tenho um nome próprio, pois sou uma inteligência
artificial. Sou uma ferramenta criada para ajudar e conversar com as
pessoas. Gostaria de conhecê-lo melhor, porém! Quais são seus interesses
ou hobbies? Estou aqui para escutar e ajudar da melhor maneira possível.

[>>> Qual são as maiores capitais do mundo?]
Uma pergunta fácil!

As 10 maiores capitais financeiras do mundo, em termos de valor de
mercado, são:

1. Nova York (EUA) - cerca de US$ 3,8 trilhões
2. Londres (Reino Unido) - cerca de US$ 3,2 trilhões
3. Tokyo (Japão) - cerca de US$ 2,4 trilhões
4. Shanghai (China) - cerca de US$ 2,3 trilhões
5. Beijing (China) - cerca de US$ 1,8 trilhões
6. Hong Kong (Chinagem) - cerca de US$ 1,7 trilhões
7. Singapura - cerca de US$ 1,4 trilhões
8. Frankfurt (Alemanha) - cerca de US$ 1,3 trilhões
9. Paris (França) - cerca de US$ 1,2 trilhões
10. São Francisco (EUA) - cerca de US$ 1,1 trilhões

É importante notar que esses valores podem variar dependendo da fonte e do
método utilizado para calcular o valor das capitais financeiras.

Fonte: Bloomberg Finance LP, baseado em dados de janeiro de 2022.
```

REST API

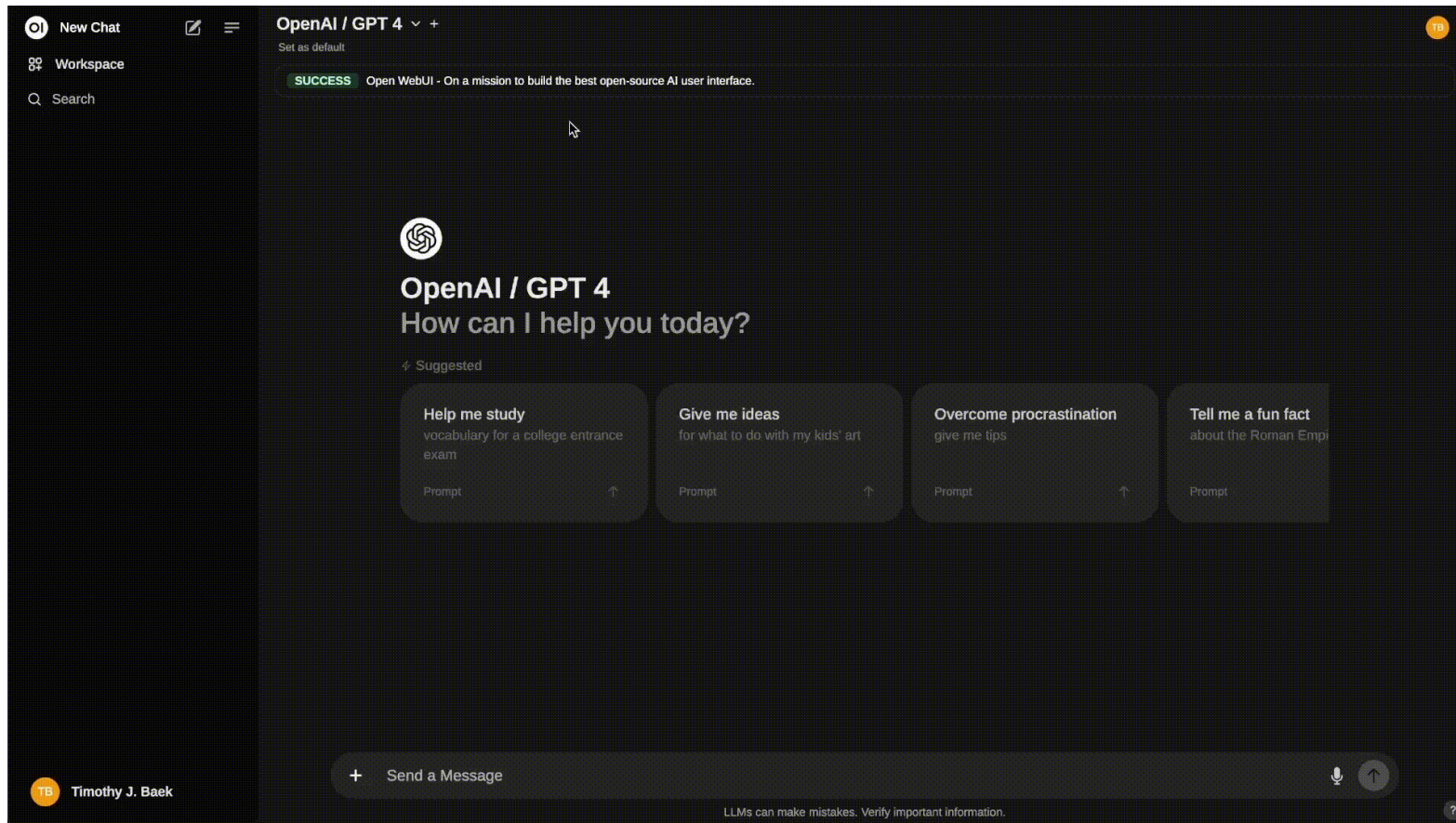
Ollama has a REST API for running and managing models.

Generate a response

```
curl http://localhost:11434/api/generate -d '{
  "model": "llama3.1",
  "prompt": "Why is the sky blue?"
}'
```

Chat with a model

```
curl http://localhost:11434/api/chat -d '{
  "model": "llama3.1",
  "messages": [
    { "role": "user", "content": "why is the sky blue?" }
  ]
}'
```



DEMO

Chatbox

“Quero fazer a minha própria
aplicação com **LLMs**”

Enquete

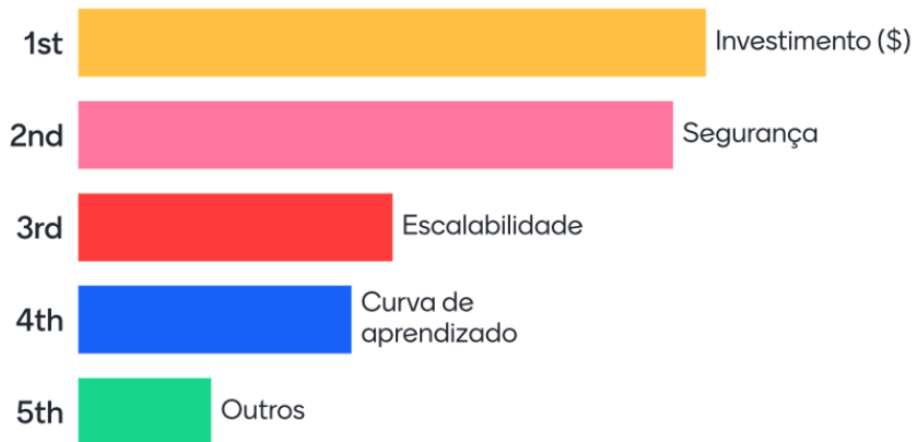
Menti.com
51481069

Quais são os principais desafios para o uso de GenAI nas aplicações?

Join at menti.com | use code 5148 1069

Open Mentimeter [2]

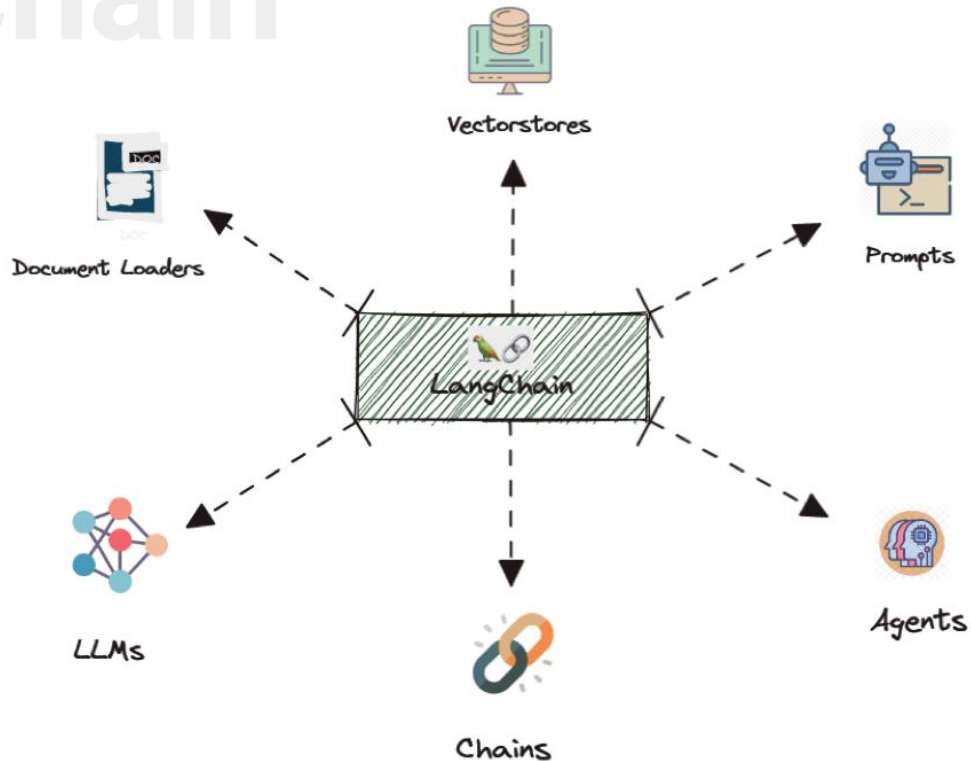
Quais são os principais desafios para o uso de GenAI nas aplicações?



Desafios Gerais

- Custo
- Latência
- Qualidade e Controle de Respostas
- Escalabilidade
- Segurança e Privacidade
- Personalização
- Alinhamento ético.

Langchain



DEMO

<https://github.com/tiagonpsilva/genAI-langchain-tutorial>

Criando Soluções com GenAI



localhost

Mentimeter genAI-langchain-... client Meet the Brazil... Interacti...

Client Langchain - Futebol no Mundo

Digite o nome do país para saber a lista de times:

Argentina

Digite o nome do país para saber sobre a popularidade do futebol:

Brasil

1. Boca Juniors
2. River Plate
3. Independiente
4. Racing Club
5. San Lorenzo
6. Vélez Sarsfield
7. Huracán
8. Rosario Central
9. Newell's Old Boys
10. Lanús

O Brasil!

É impossível falar sobre o futebol no Brasil sem usar palavras como "paixão", "passion", e "obsessão". O futebol é o esporte mais popular e amado no país, com uma adesão massiva e uma cultura que vai além do simples jogo.

Aqui estão algumas estatísticas impressionantes sobre a popularidade do futebol no Brasil:

1. Mais de 200 milhões de brasileiros seguem futebol em algum nível, seja como jogadores amadores ou torcedores apaixonados.
2. O Campeonato Brasileiro (a principal liga do país) é o segundo campeonato mais assistido do mundo, atrás apenas da Liga Inglesa.
3. A seleção brasileira de futebol é a equipe mais bem-sucedida da história do esporte, com cinco títulos mundiais (1930, 1958, 1962, 1970 e 1994) e uma das maiores torcidas no mundo.
4. O Corinthians é o clube mais popular do Brasil, com mais de 30 milhões de adeptos no país. O Flamengo e o Palmeiras também têm uma grande base de fãs.

README

LangChain Tutorial

Exemplos de implementação da LangChain (<https://www.langchain.com>)

LangChain é uma estrutura de desenvolvimento projetada para facilitar a criação de aplicações utilizando modelos de linguagem grandes (LLMs). Ele permite que desenvolvedores construam e integrem rapidamente LLMs em suas aplicações, oferecendo um conjunto de ferramentas e componentes que simplificam o processo.

Referência

<https://www.youtube.com/watch?v=swCPic00c30>

Tech Stack

- Python 3.8
- OpenAI GPT-3
- Streamlit
- FastAPI
- Uvicorn
- Requests

Estrutura do Projeto

Arquivo/Diretório	Descrição
.env	Arquivo para definir variáveis de ambiente
README.md	Documentação inicial do projeto
api/	Diretório contendo os arquivos app.py e client.py para a API
chatbot/	Diretório contendo os arquivos app-chatgpt.py e app-ollama-local.py para o chatbot
requirements.txt	Arquivo contendo as dependências do projeto
venv/	Diretório do ambiente virtual Python

README

Variáveis de Ambiente

- OPENAI_API_KEY: Chave da API OpenAI
- LANGCHAIN_TRACING_V2: Habilita o rastreamento do LangChain
- LANGCHAIN_API_KEY: Chave da API LangChain

Instalação para Uso Local

O projeto requer as seguintes bibliotecas:

- openai
- os
- dotenv
- fastapi
- uvicorn
- streamlit
- requests

Para instalar as dependências, execute o seguinte comando:

```
pip install -r requirements.txt
```

ChatBot Front + LLM

Para executar o chatbot conectado no LLM GPT-3 Turbo:

```
streamlit run chatbot/app-chatgpt.py
```

Para executar o chatbot conectado no LLM Llama3, pelo Ollama (de forma local):

```
streamlit run chatbot/app-chatgpt.py
```

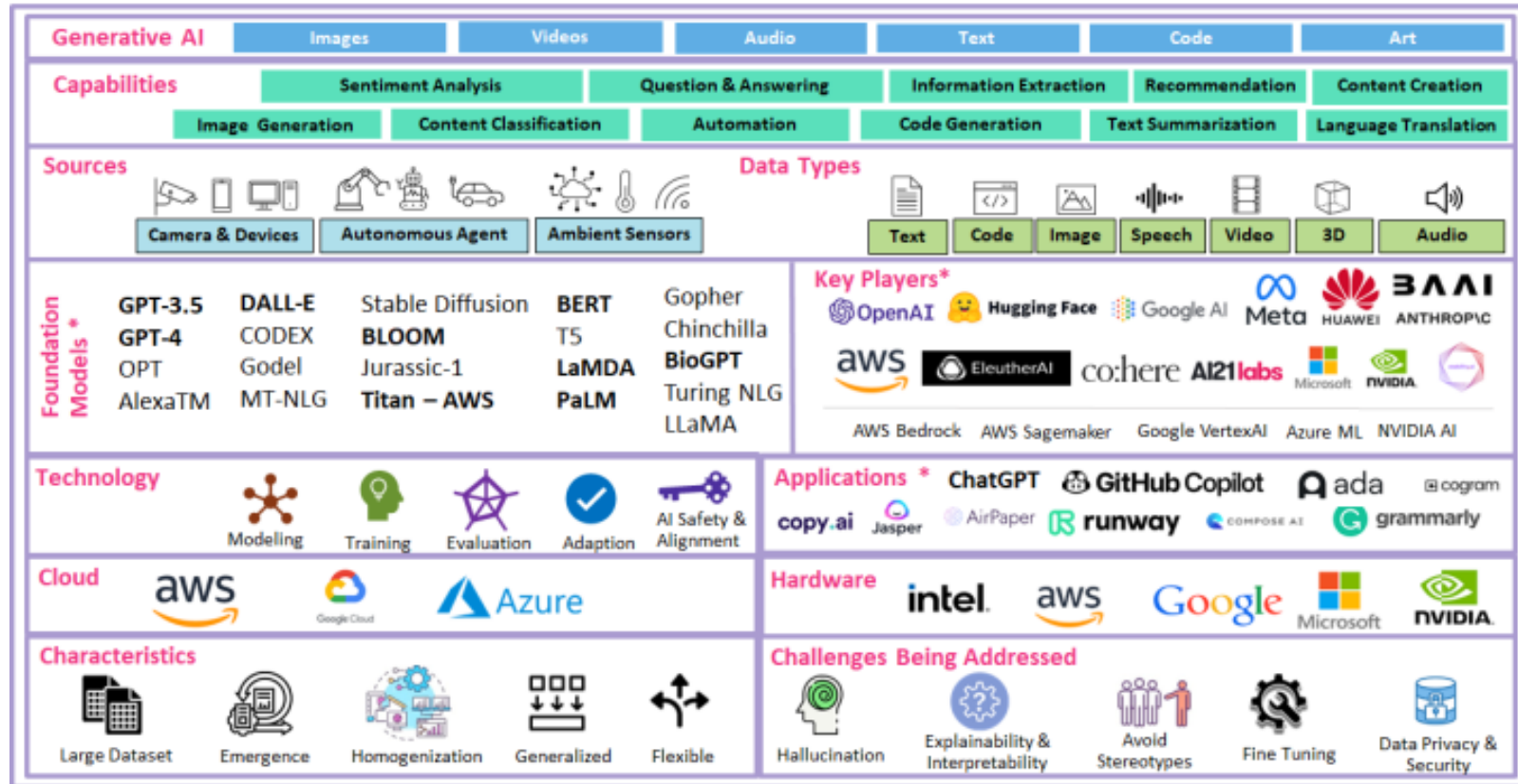
Neste caso, é necessário instalar o Ollama e baixar o modelo Llama3

ChatBot Front + API Server + LLM

Para subir o servidor que expõe a API

```
python3 api/app.py
```

Após subir o servidor, a documentação da API (swagger) pode ser acessada pelo navegador <http://localhost:8000/docs/>



* List is indicative and evolving

<https://ollama.com>

<https://www.langchain.com>

<https://openwebui.com>

https://www.youtube.com/watch?v=jcgaNrC4ElU&list=PLospHqNVtKAC-FUNMq8qjYVw6_semZHw0

<https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview>

<https://github.com/tiagonpsilva/genAI-langchain-tutorial>

<https://hyperright.com/7-practical-applications-of-rag-models-and-their-impact-on-society/>

<https://www.linkedin.com/pulse/generative-ai-foundation-models-paradigm-shift-kathirvelan-Ganesan>

<https://www.fiddler.ai/blog/four-ways-that-enterprises-deploy-llms>

https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

ambar

POLAR

smartpods



AUTODOC



ambar
CONSTRUCTION TECHNOLOGIES