elastic

Workshop Criando chatbots para conversar com seus dados privados através da IA, LLMs e Elasticsearch

Alex Salgado
Senior Developer Advocate, Elastic

@alexsalgadoprof - redes sociais - TDC Brasília

Alex Salgado

Senior Developer Advocate LATAM

@alexsalgadoprof

salgado

@alexsalgadoprof

/in/alex-salgado/

- **Mestre** em Ciência da Computação pela UFF (Games)
- **MBA** UFF
- **PhD** Candidate UFF: Robótica/Visão Computacional

- **+ 25 anos** de experiência na área de desenvolvimento de software
- Ocupei diversos cargos, trabalhando em **startups**, pequenas e grandes empresas como Oracle, CSN, BRQ/IBM, **Chemtech/Siemens (9 anos)**.
- **8 anos** como professor universitário

aws certified Solutions Architect Associate

aws certified Cloud Practitioner

LEVEL 2 CERTIFIED Educator Google for Education

Microsoft CERTIFIED

Three solutions powered by one stack

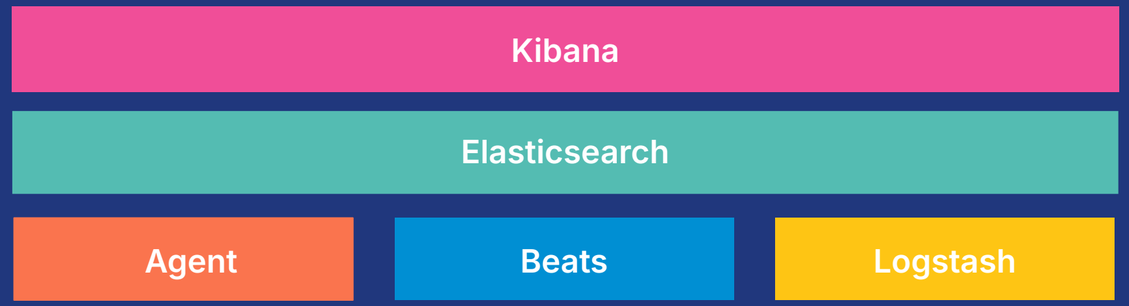
3 solutions

Enterprise Search

Observability

Security

Powered by the Elastic Stack



Deployed anywhere

Elastic Cloud

Elastic Cloud Enterprise

Elastic Cloud on Kubernetes

Saas

Orchestration

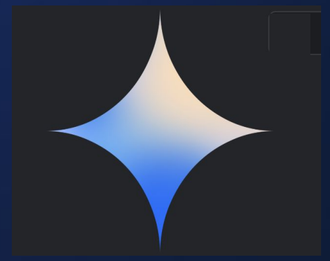
elastic

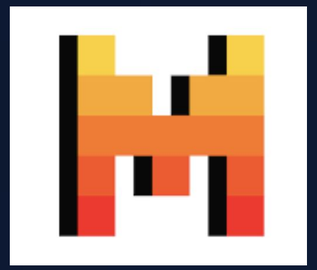
Trusted by Organizations Around the World

TECHNOLOGY	FINANCE	TELCO	CONSUMER	HEALTHCARE	PUBLIC SECTOR	AUTOMOTIVE / TRANSPORTATION	RETAIL
Adobe	BARCLAYS	orange	Uber	VITAS Healthcare	Lawrence Livermore National Laboratory	VOLVO Volvo Group	AutoZone
CISCO	ZURICH	dish media	Grab	UCLA Health	OAK RIDGE National Laboratory	Audi	The Home Depot
workday	USAA	COMCAST	Miles & More Lufthansa	Yale NewHaven Health	De Watergroep WATER. VANDAAG EN MORGEN.	JAGUAR LAND ROVER	ebay
Microsoft	Swift	verizon	ACTIVISION BLIZZARD	MAYO CLINIC	JPL Jet Propulsion Laboratory	BMW	Kroger
INGRAM MICRO	Postbank	T-Mobile	lyft	Pfizer	MENTAT COMPUTE OPTIMIZATION	Volkswagen	Walgreens

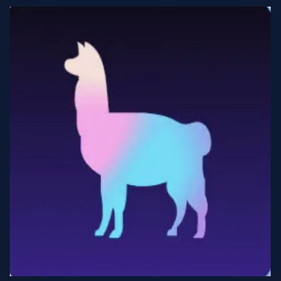


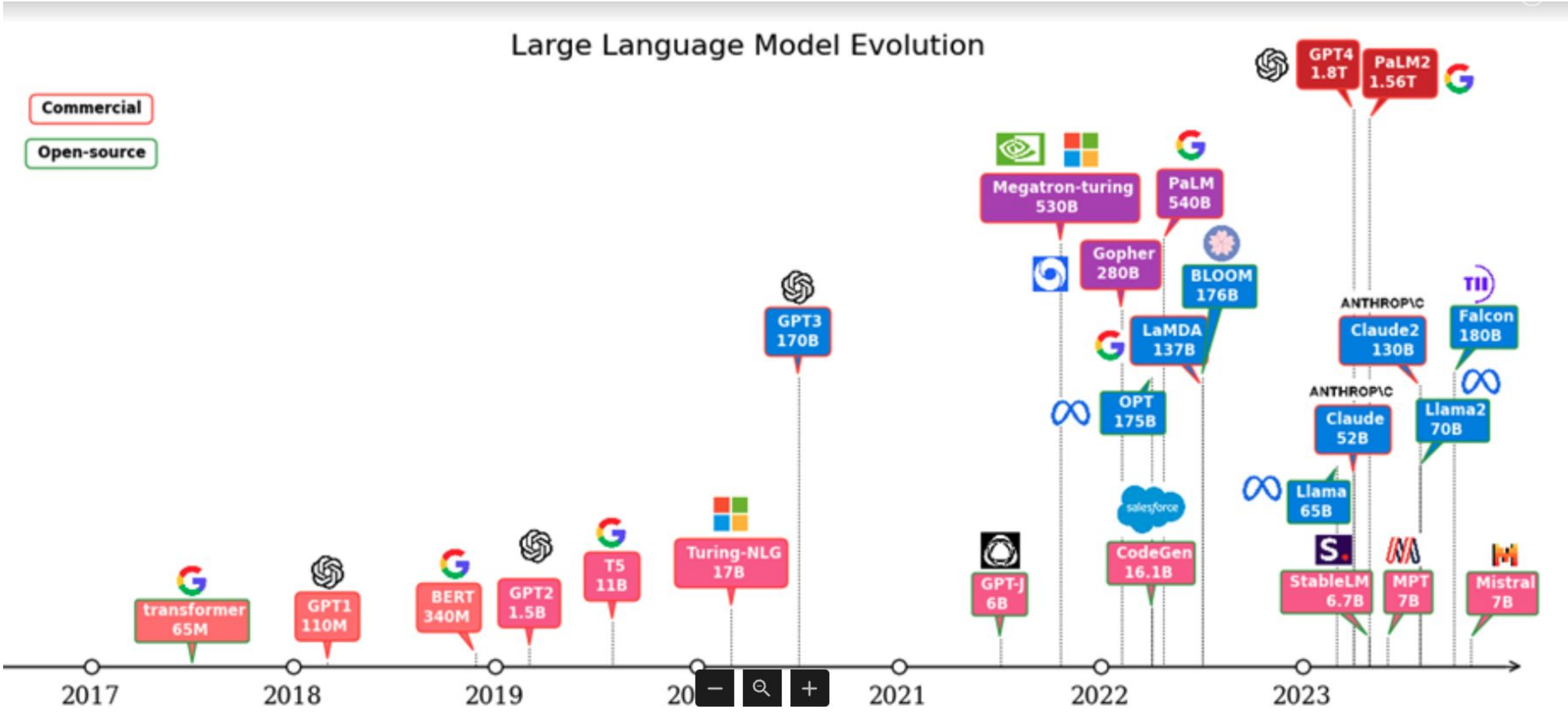




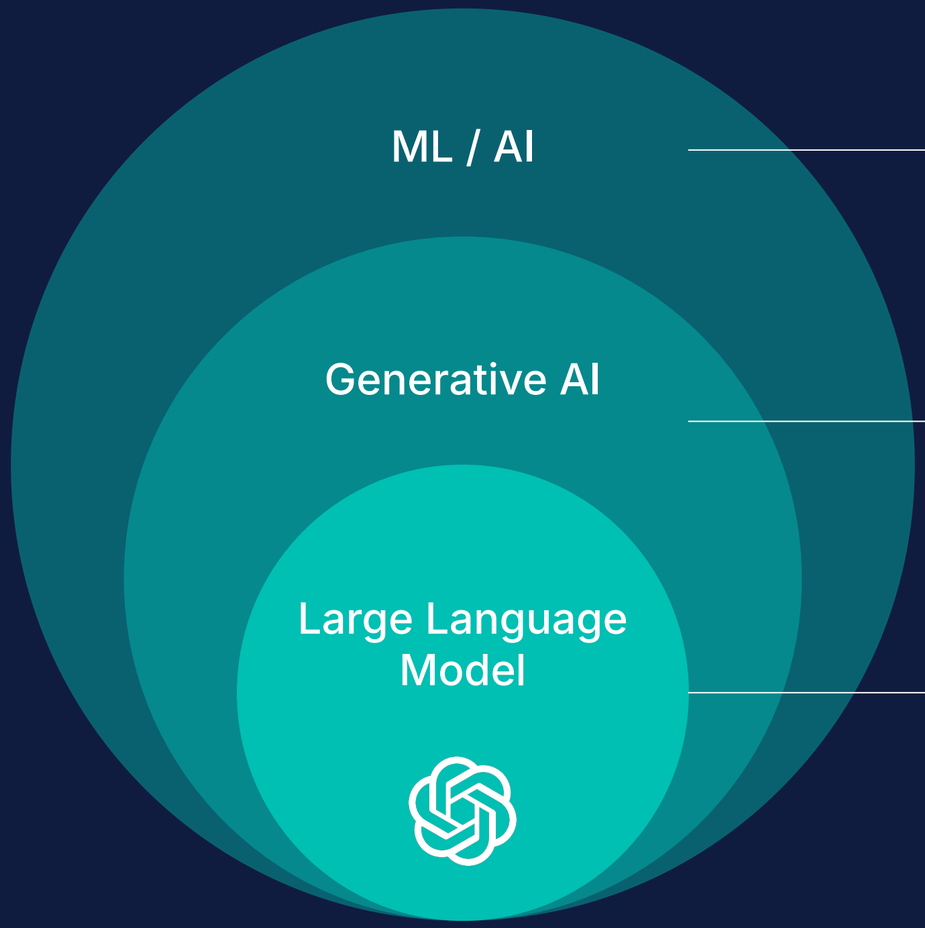








Large Language Model Evolution
Commercial
Open-source
transformer 65M
GPT1 110M
BERT 340M
GPT2 1.5B
T5 11B
Turing-NLG 17B
GPT3 170B
GPT-J 6B
Megatron-turing 530B
Gopher 280B
OPT 175B
PaLM 540B
LaMDA 137B
BLOOM 176B
CodeGen 16.1B
GPT4 1.8T
PaLM2 1.56T
Claude2 130B
Claude 52B
Llama 65B
StableLM 6.7B
Falcon 180B
Llama2 70B
MPT 7B
Mistral 7B
2017
2018
2019
2020
2021
2022
2023

Aplicações interessantes de ML, IA generativa e LLMs



	O que é ?	Aplicações
ML / AI	Algoritmos programados para fazer previsões com base em dados	Reconhecimento de imagem, processamento de linguagem natural, reconhecimento de fala
Generative AI	Algoritmos de IA projetados para criar novos dados	Chatbots, geradores de texto, geradores de imagem, geradores de música
Large Language Model	Algoritmos de aprendizado profundo que podem gerar texto	Geradores de texto, tradução, escrita, resposta a perguntas

elastic

What is a Vector?

elastic

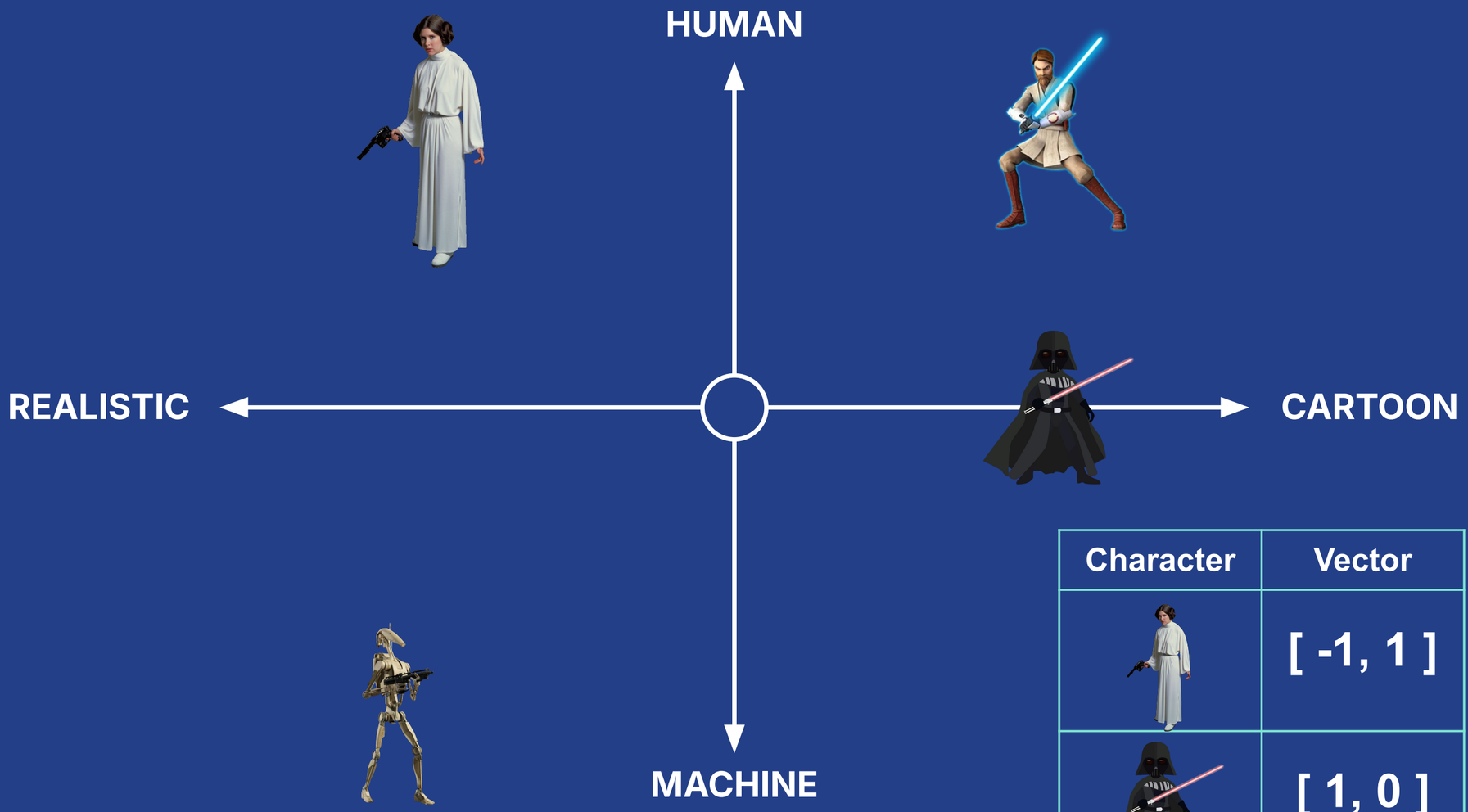
Embeddings represent your data

Example: 1-dimensional vector

Character	Vector
	[-1]
	[1]

elastic

Multiple dimensions represent different data aspects



Character	Vector
	[-1, 1]
	[1, 0]

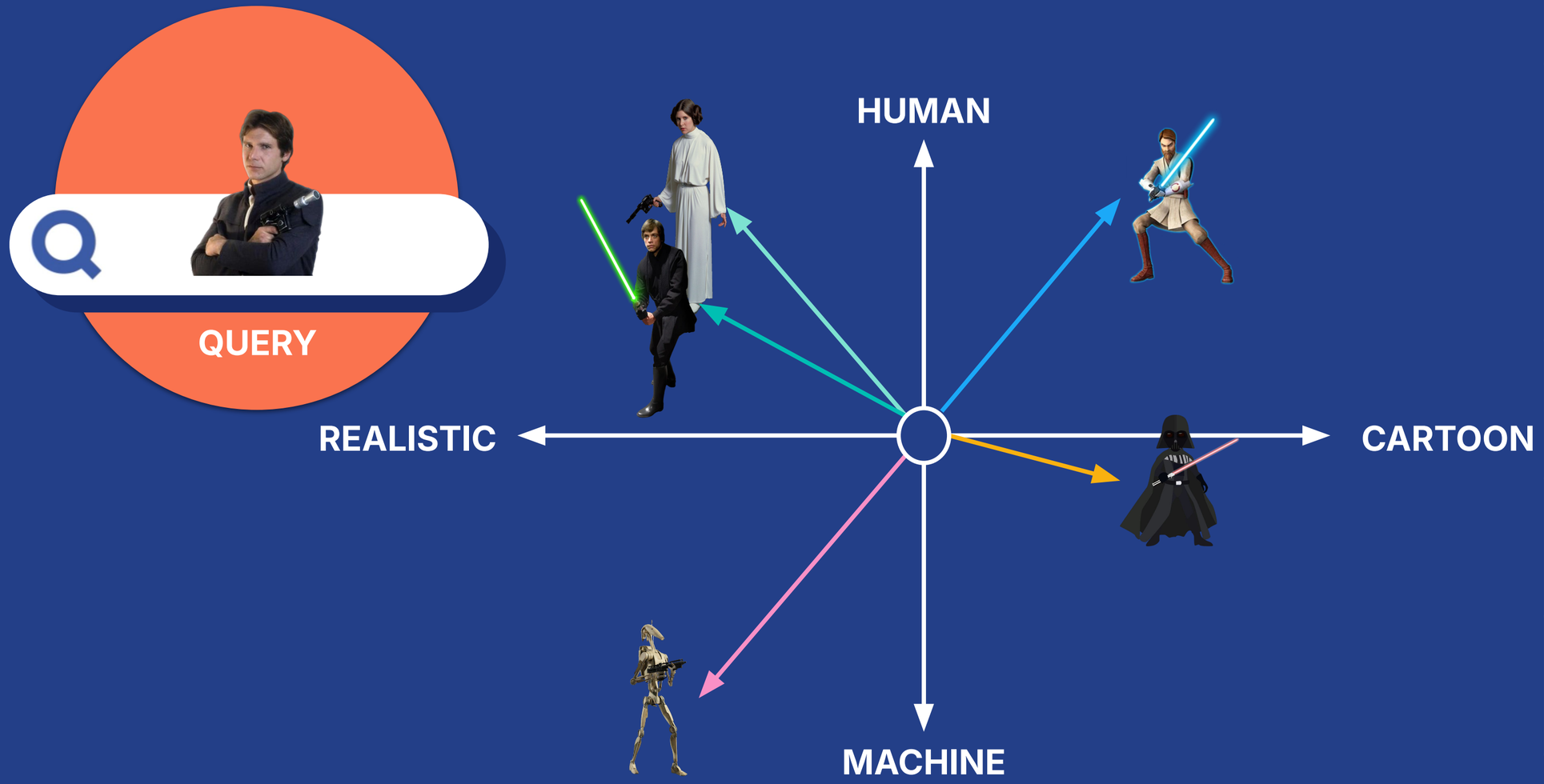
elastic

Similar data is grouped together

Character	Vector
	[-1.0, 1.0]
	[1.0, -0.1]
	[-1.0, 0.8]

elastic

Vector search ranks objects by similarity (relevance) to the query

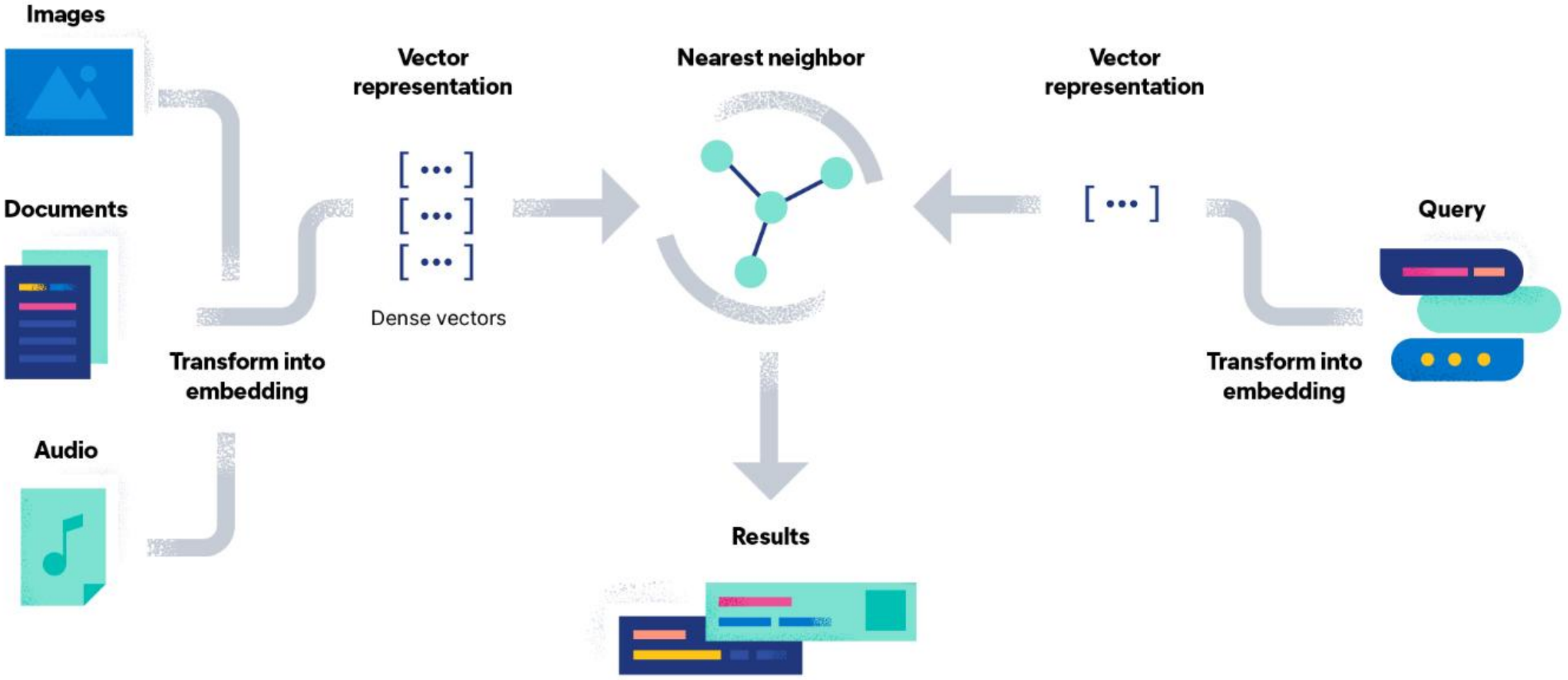


Relevance	Result
Query	
1	
2	
3	
4	
5	

elastic

Vector search conceptual architecture

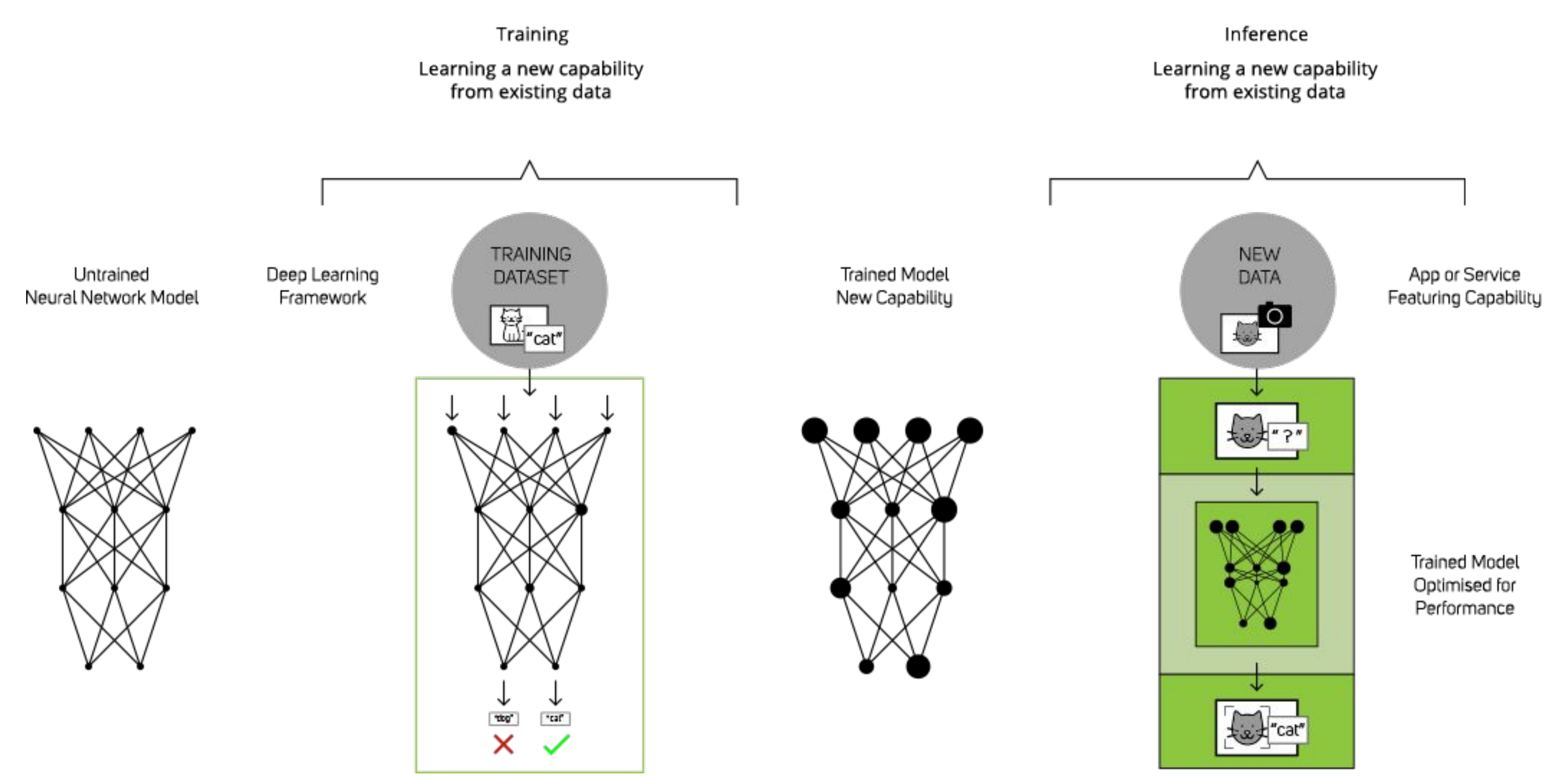
Use vector nearest neighbor to generate a search ranking



elastic

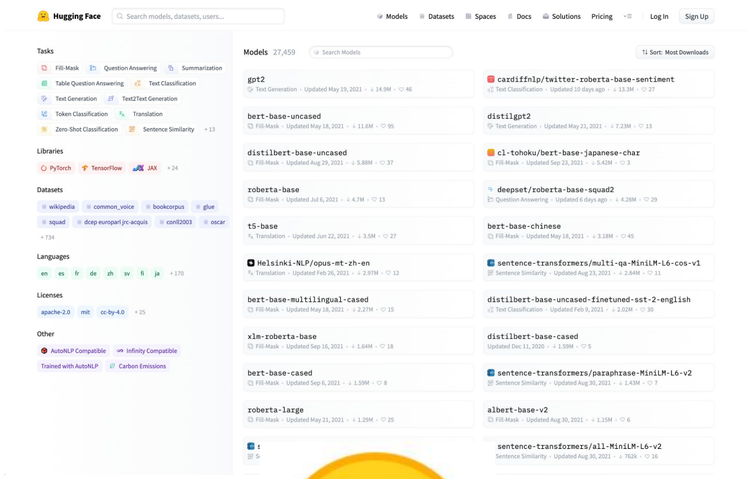
What is a Model?

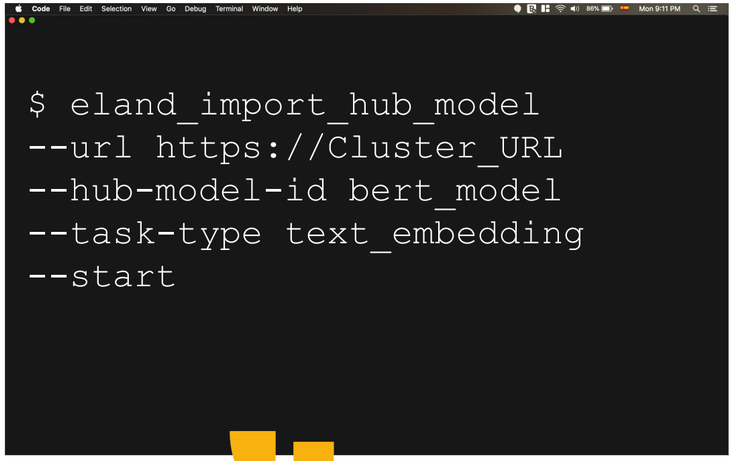
elastic

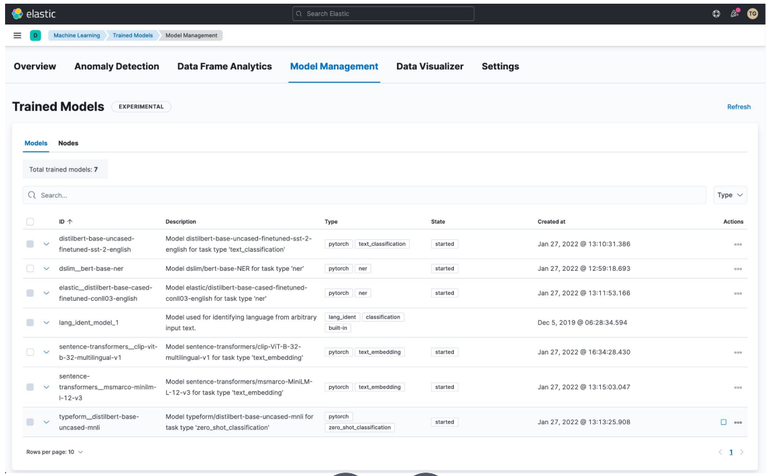


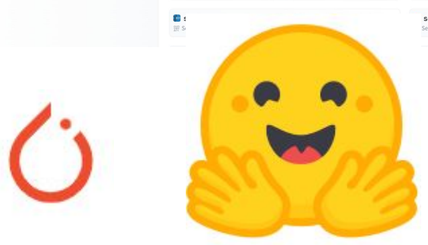
https://www.scan.co.uk/business/deep-learning-guide

Eland Imports PyTorch Models

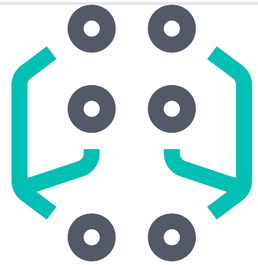








Huggingface.co



Inference, not training

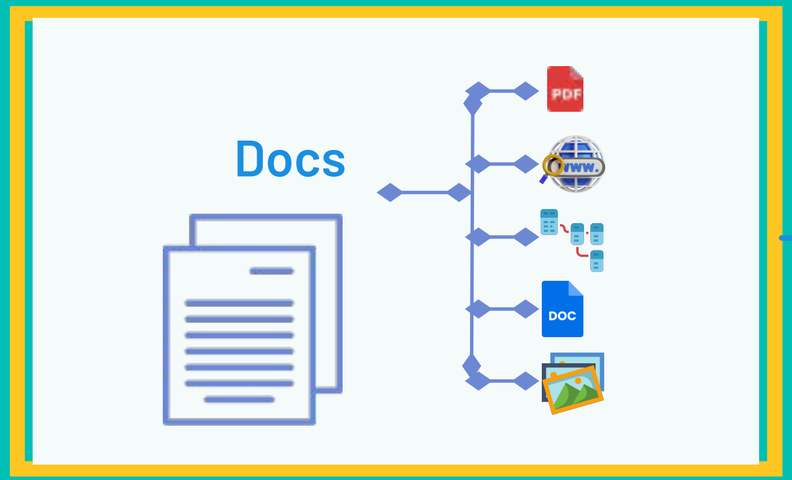
elastic

Chatbot Pipeline

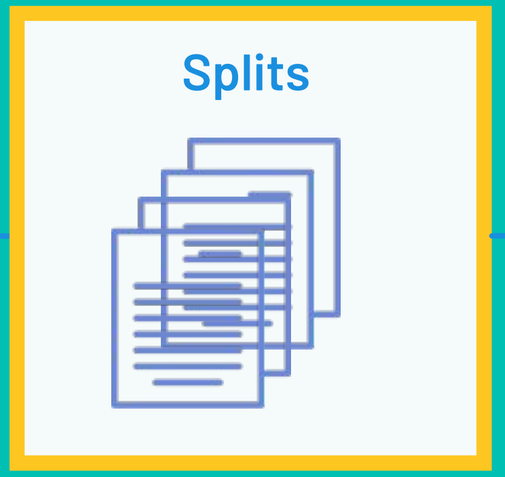
Converse com seus dados com este pipeline

Carregamento do banco vetorial

Carregamento de documentos



divisão/splitting



armazenaemnto/storage

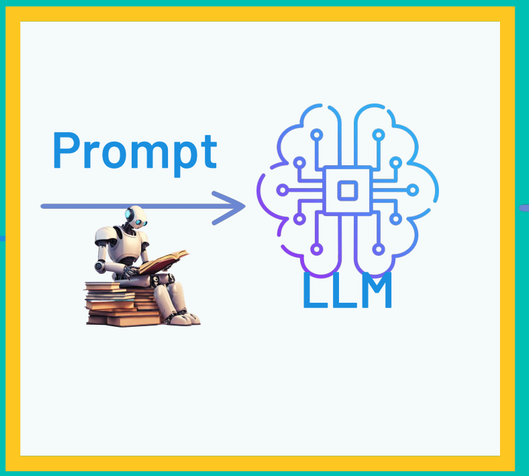
Recuperação/ Retrieval

query

Storage

Retrieval

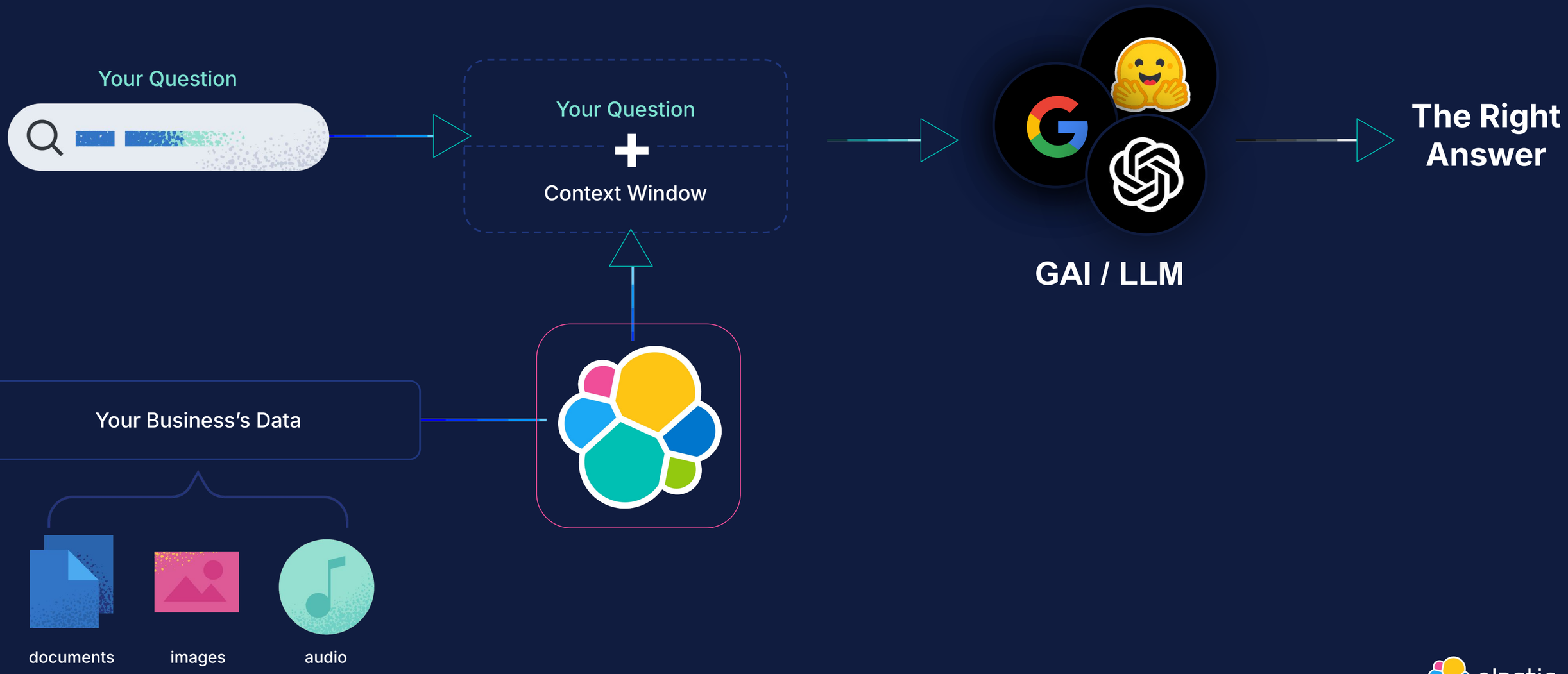
Output



resposta

rr

RAG



elastic

rr

Chatbot→RAG

Chat iteration

Your Question

Your Question
+
Context Window
+
Memory

GAI / LLM

The Right Answer

Your Business's Data

documents

images

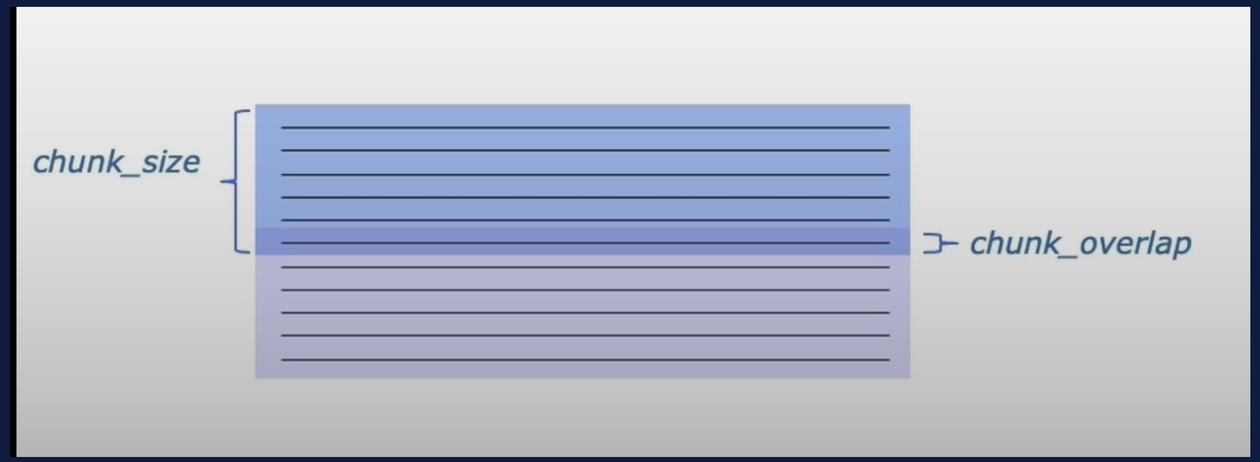
audio

elastic

Exemplo de Splitter

```
langchain.text_splitter.CharacterTextSplitter
(
    separator: str = "\n\n"
    chunk_size=4000,
    chunk_overlap=200,
    length_function=<builtin function len>,
)
```

Methods:
creeate_documents() - Create documents from a list of texts.
s**plit_documents()** - Split documents.



elastic

Tipos de Splitters

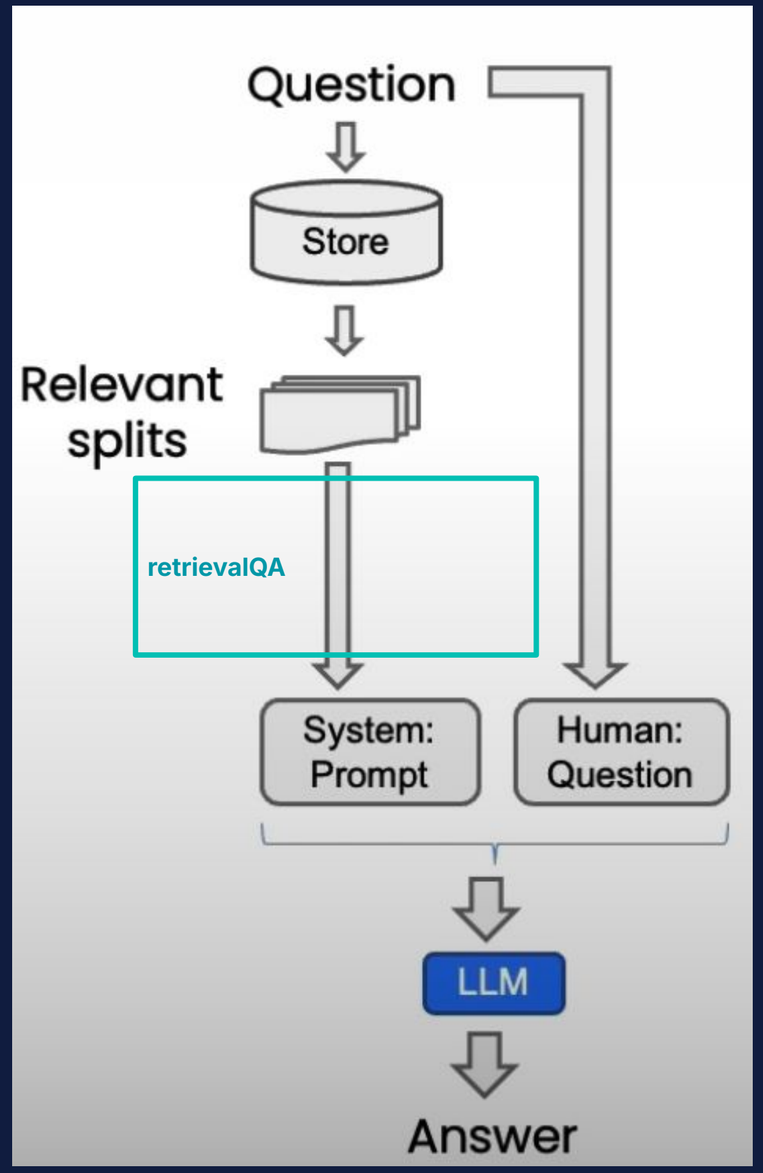
Langchain.text_splitter.

- **CharacterTextSplitter()** - Implementation of splitting text that looks at characters.
- **MarkdownHeaderTextSplitter()** - Implementation of splitting markdown files based on specified headers.
- **TokenTextSplitter()** - Implementation of splitting text that looks at tokens.
- **SentenceTransformersTokenTextSplitter()** - Implementation of splitting text that looks at tokens.
- **RecursiveCharacterTextSplitter()** - Implementation of splitting text that looks at characters. Recursively tries to split by different characters to find one that works.
- **Language() –** for CPP, Python, Ruby, Markdown etc
- **NLTKTextSplitter()** - Implementation of splitting text that looks at sentences using NLTK (Natural Language Tool Kit)
- **SpacyTextSplitter()** - Implementation of splitting text that looks at sentences using Spacy

Question Answering

RetrievalQA chain

```
RetrievalQA.from_chain_type(, chain_type="stuff",...)
```



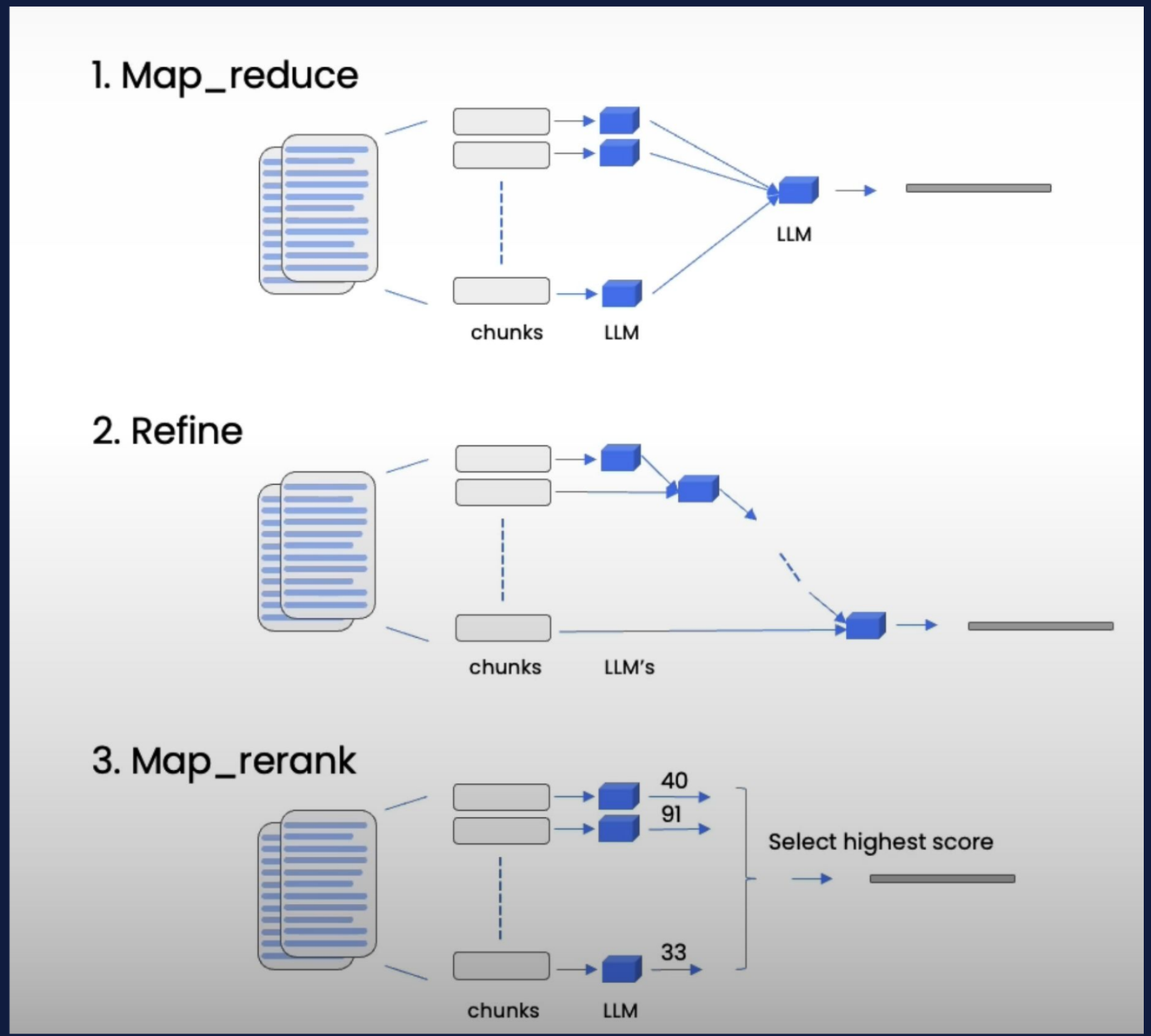
A pergunta é aplicada ao Vector Store como uma consulta

Vector Store fornece **k** documentos relevantes

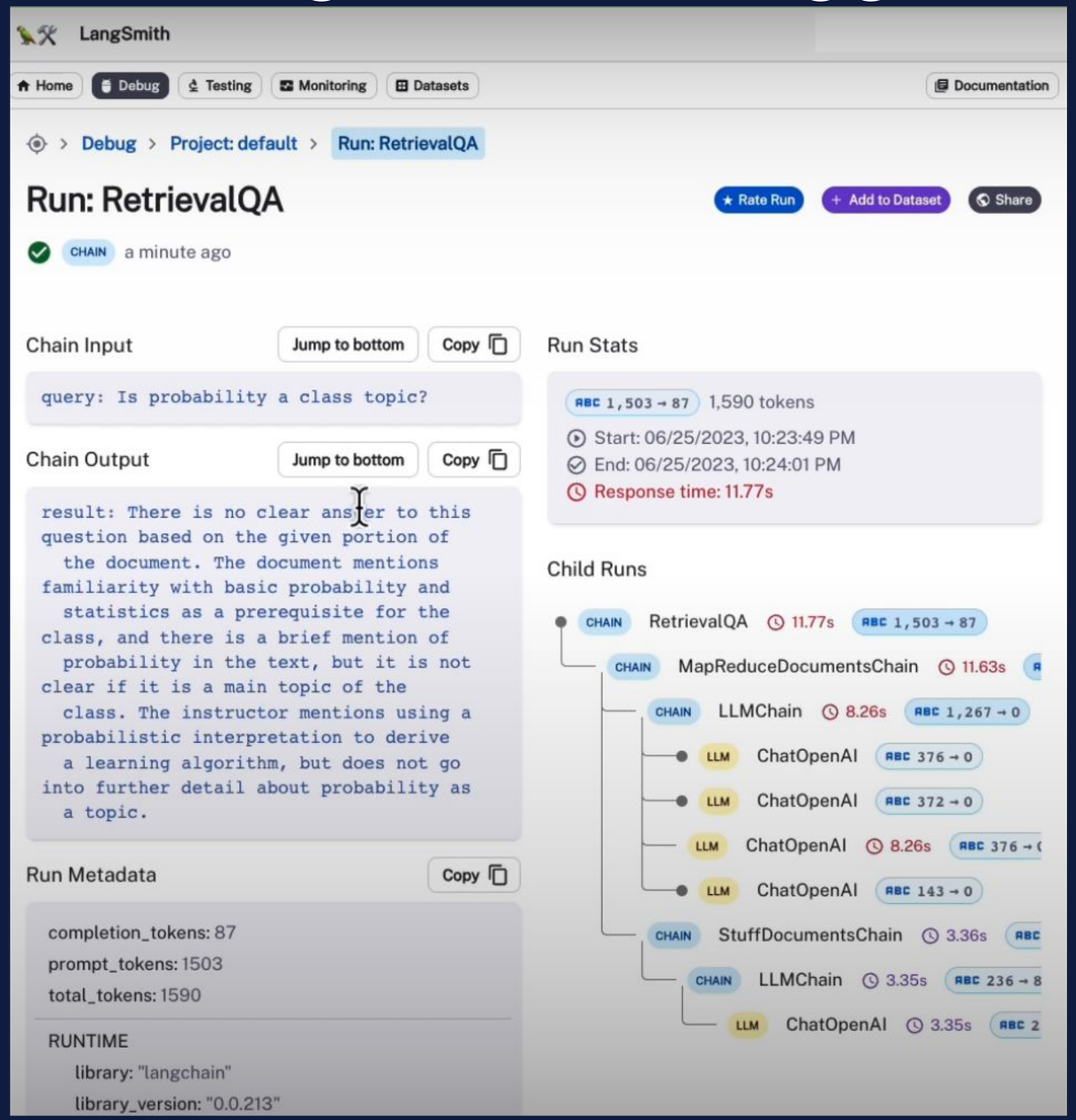
Documentos e a pergunta original são enviados para um LLM

elastic

Tipos de Question Answerings



LangSmith Debugger



elastic

Chat

Conversational Memory

LLMs are stateless

Com memória conversacional

"Estou interessado em integrar LLMs com conhecimento externo."

LLMs são ótimos em gerar textos semelhantes aos humanos. No entanto, integrar conhecimento externo pode aprimorar ainda mais suas capacidades.

"Quais são os diferentes métodos possíveis para fazer isso?"

Você poderia usar grafos de conhecimento pré-existentes, permitir que LLMs acessem ferramentas como APIs, ou fazer augmentação de recuperação com bancos de dados vetoriais!

Histórico de Conversa

Interessante! O que era mesmo que eu queria saber de novo?

Você estava interessado em integrar LLMs com conhecimento externo.

Sem memória conversacional

"Não há histórico de conversa armazenado"

Histórico de Conversa

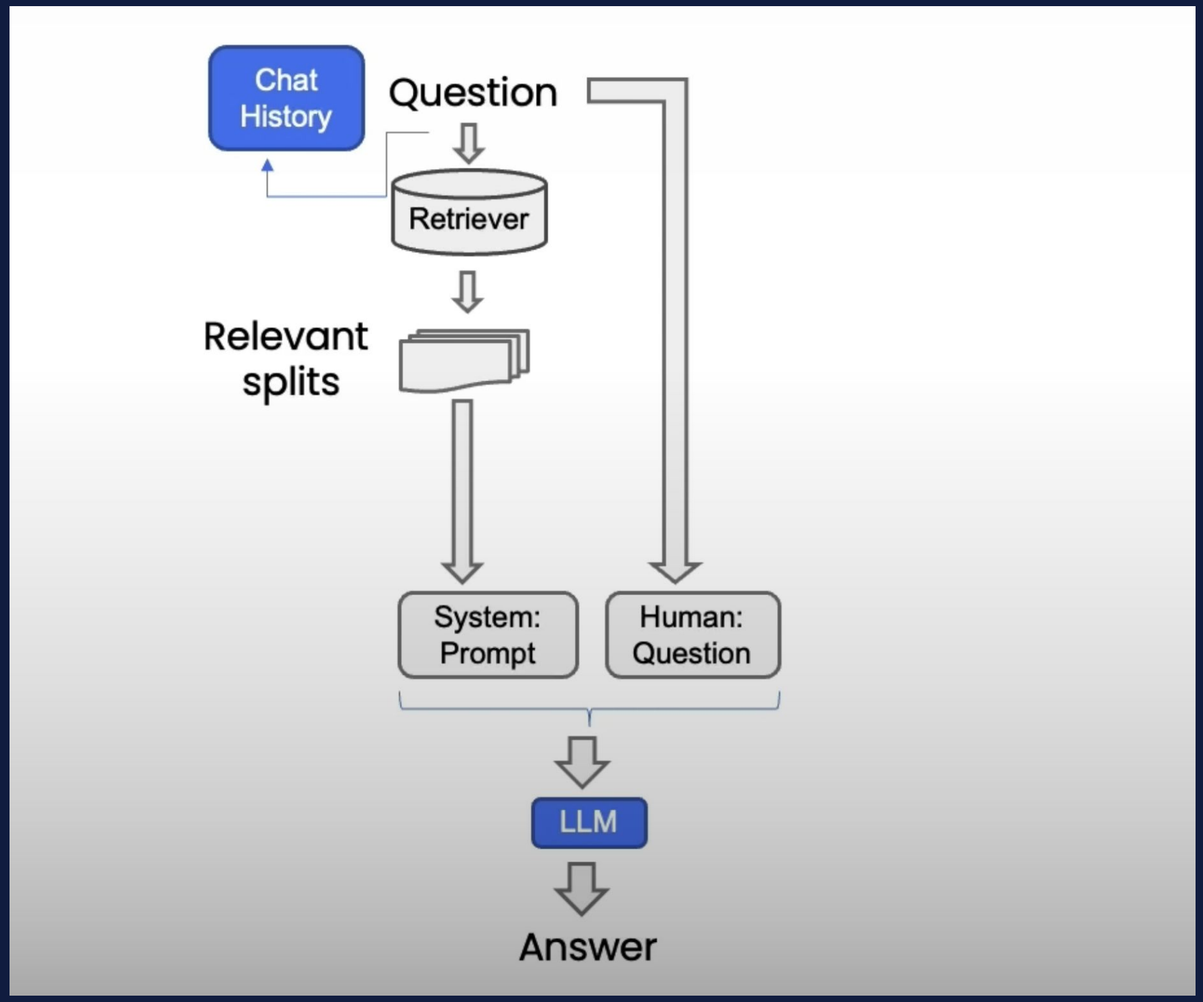
Interessante! O que era mesmo que eu queria saber de novo?

Desculpe, não faço ideia do que você está falando!

elastic

ConversationRetrievalChain

```
qa = ConversationalRetrievalChain.from_llm(ChatOpenAI(temperature=0),
vectorstore.as_retriever(), memory=memory)
```



elastic

Demo

Próximos passos

Referências

https://github.com/elastic/elasticsearch-labs/blob/main/notebooks/integrations/gemini/qa-langchain-gemini-elasticsearch.ipynb

https://github.com/salgado/meetup_goiania/blob/main/Meetup_Goiania_qa_langchain_gemini_elasticsearch.ipynb

https://www.langchain.com/

https://www.deeplearning.ai/

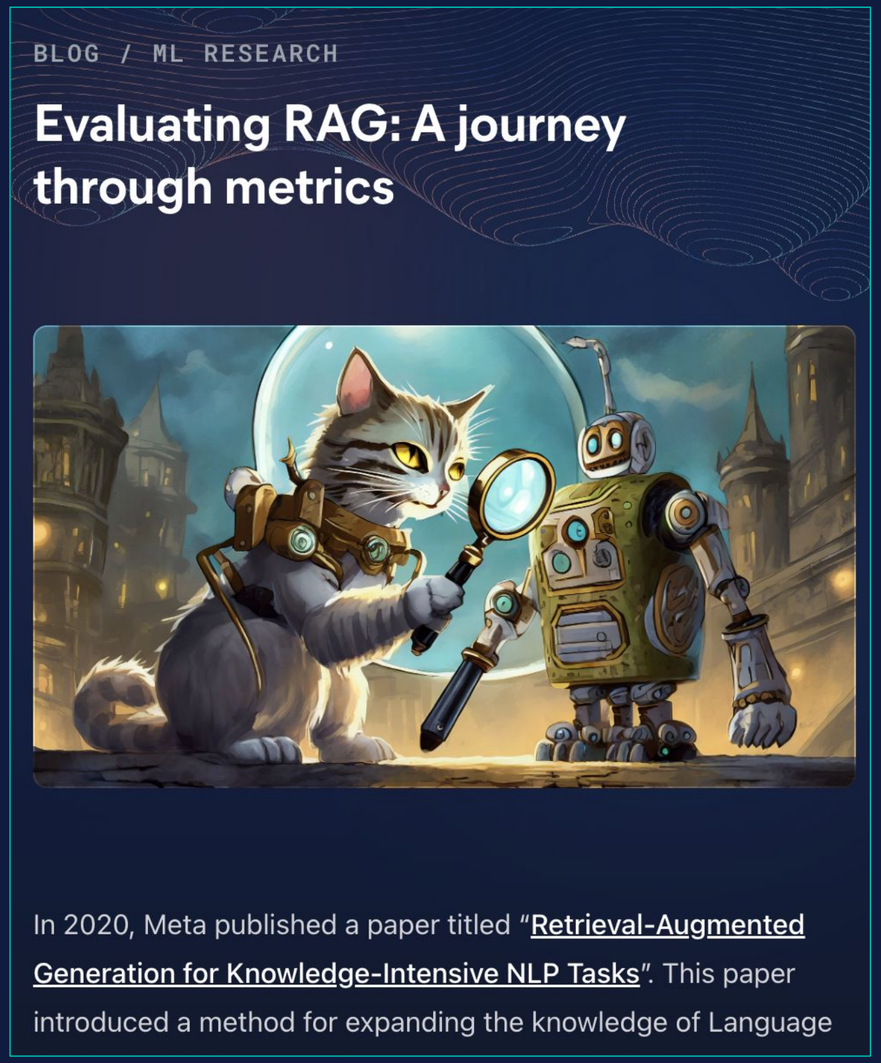
https://www.elastic.co/search-labs/tutorials/chatbot-tutorial/welcome

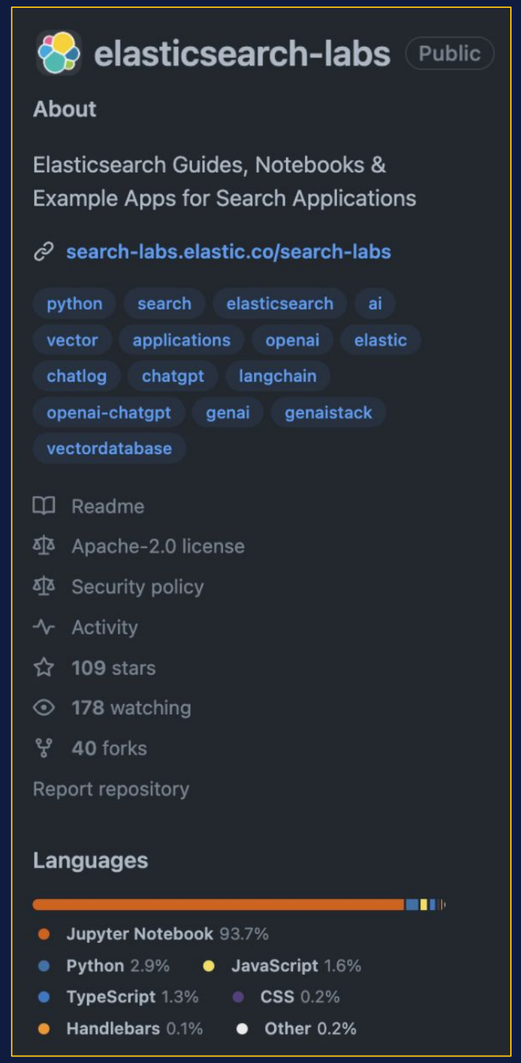
elastic

Recursos para desenvolvedores: **Elasticsearch Labs**

elastic.co/**search-labs**

github.com/elastic/**elasticsearch-labs**





Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

elastic

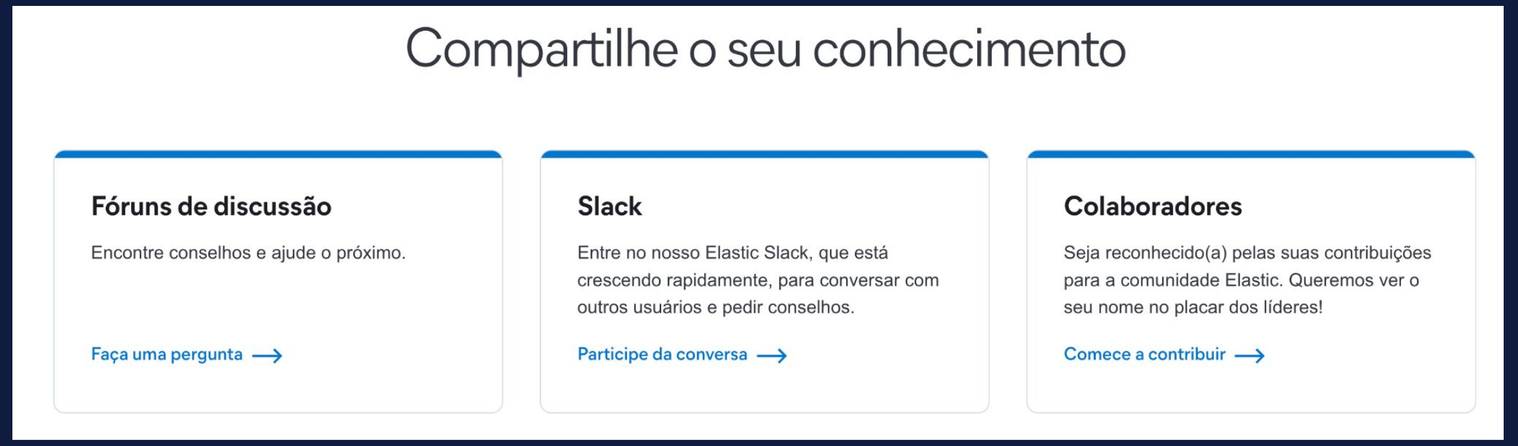
Recursos para desenvolvedores:
Junte-se à Comunidade Elastic

Elastic User Groups

Estamos sempre em busca de organizadores, palestrantes e participantes.
Encontre mais eventos Elastic em todo o mundo em **community.elastic.co**

Envie-nos um e-mail para
meetups@elastic.co se você tem interesse!

elastic.co/**community**



elastic

Obrigado

Alex Salgado
Senior Developer Advocate, Elastic

@alexsalgadopro

salgado

@alexsalgadoprof

/in/alex-salgado/