



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



Three solutions powered by one stack

3 solutions



Enterprise Search

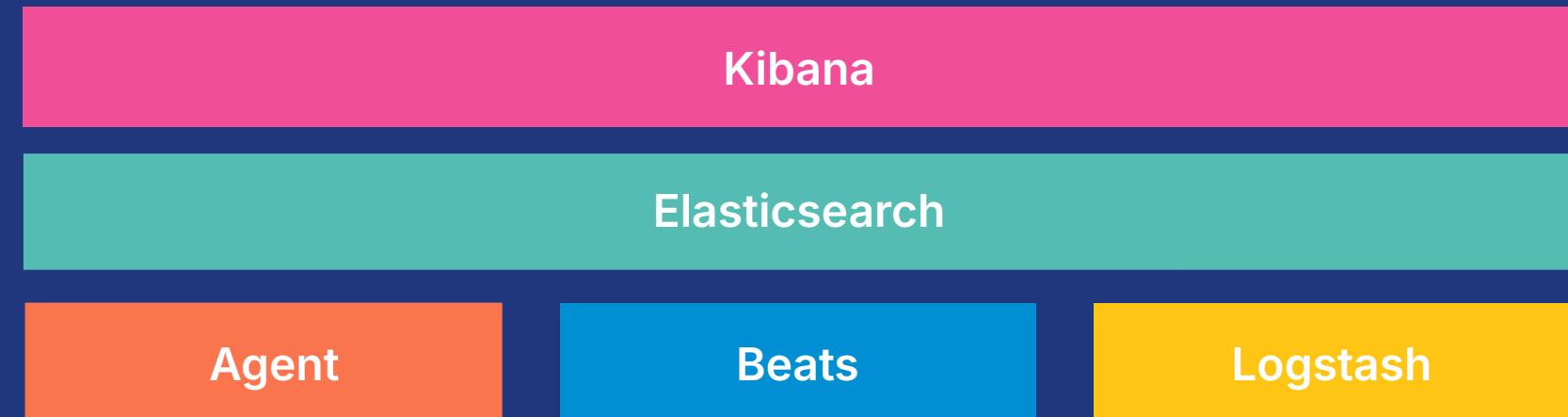


Observability



Security

Powered by
the Elastic Stack



Deployed
anywhere



Elastic Cloud



Elastic Cloud
Enterprise



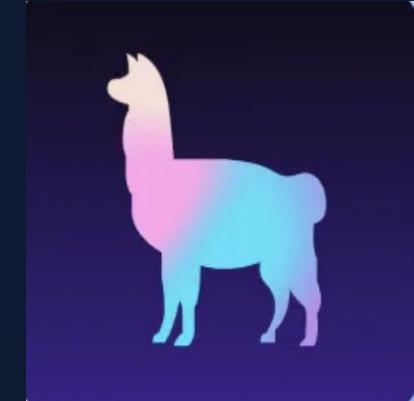
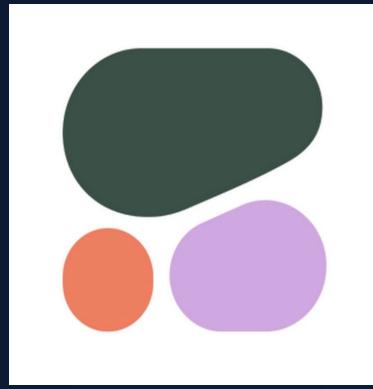
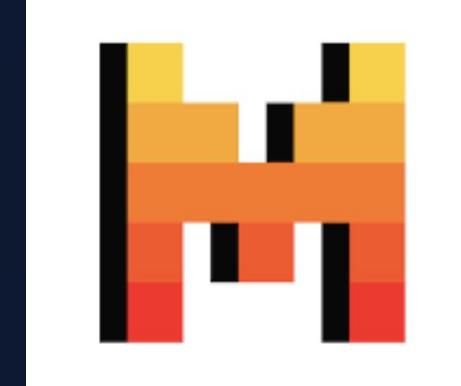
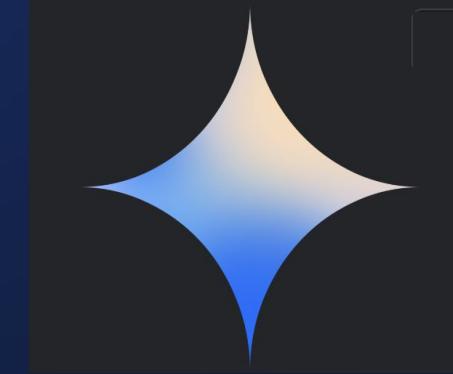
Elastic Cloud
on Kubernetes

SaaS

Orchestration

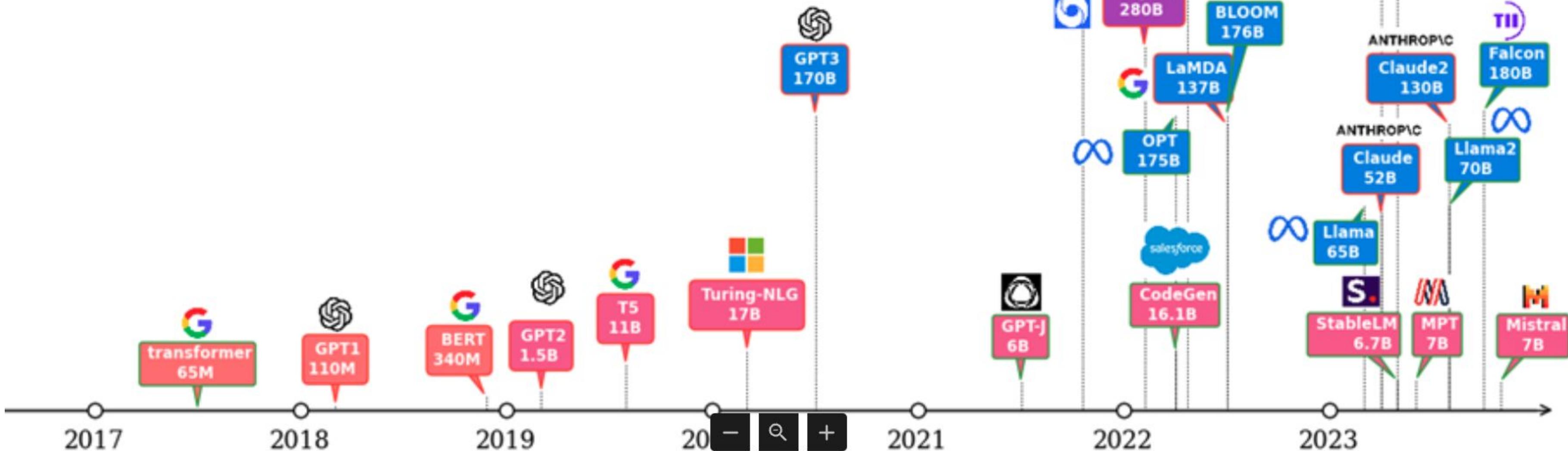
Trusted by **Organizations** Around the World

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 <small>Lufthansa</small>	 Yale New Haven Health	 De Watergroep <small>WATER. VANDAAG EN MORGEN.</small>	 JAGUAR  LAND ROVER	 eBay™
 Microsoft	 Swift	 verizon	 ACTIVISION BLIZZARD	 MAYO CLINIC	 JPL <small>Jet Propulsion Laboratory</small>	 BMW	 Kroger
 INGRAM MICRO®	 Postbank	 T-Mobile™	 lyft	 Pfizer	 MENTAT™ <small>COMPUTE OPTIMIZATION</small>	 VW	 Walgreens

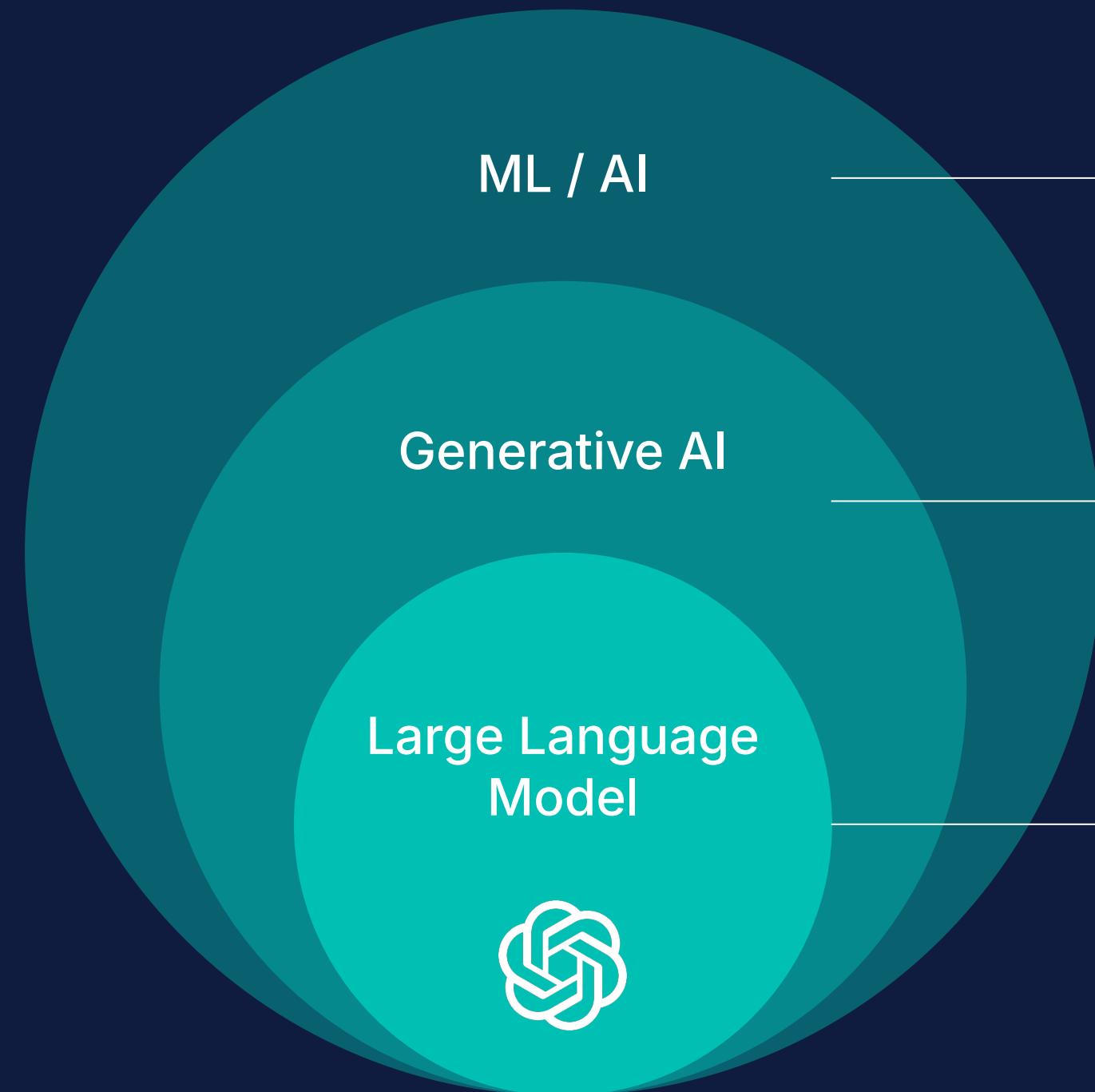


Large Language Model Evolution

Commercial
Open-source



Aplicações interessantes de ML, IA generativa e LLMs



O que é ?

Algoritmos programados para fazer previsões com base em dados

Algoritmos de IA projetados para criar novos dados

Algoritmos de aprendizado profundo que podem gerar texto

Aplicações

Reconhecimento de imagem, processamento de linguagem natural, reconhecimento de fala

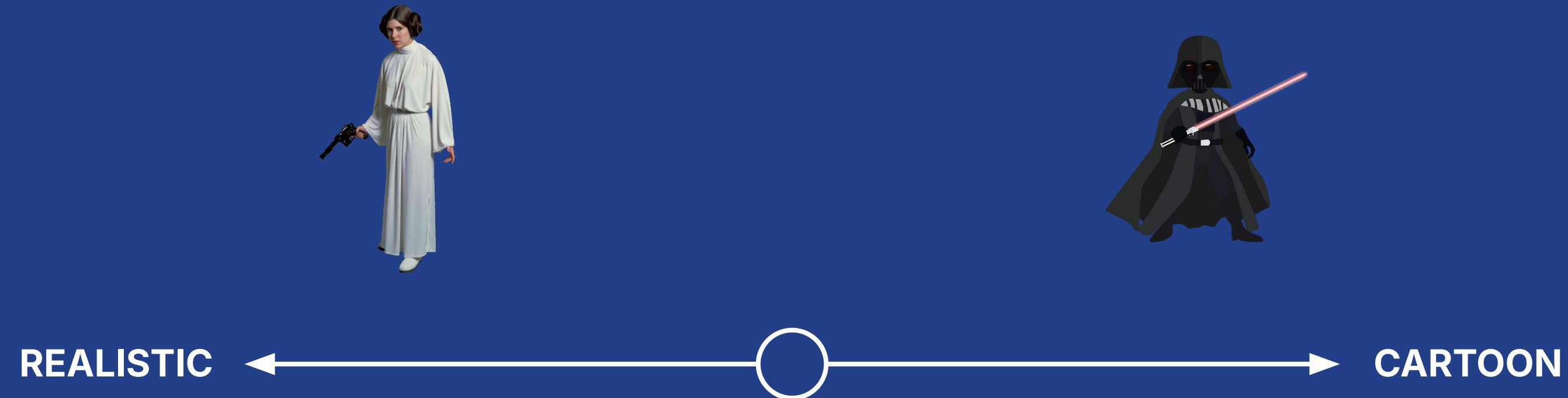
Chatbots, geradores de texto, geradores de imagem, geradores de música

Geradores de texto, tradução, escrita, resposta a perguntas

What is a Vector?

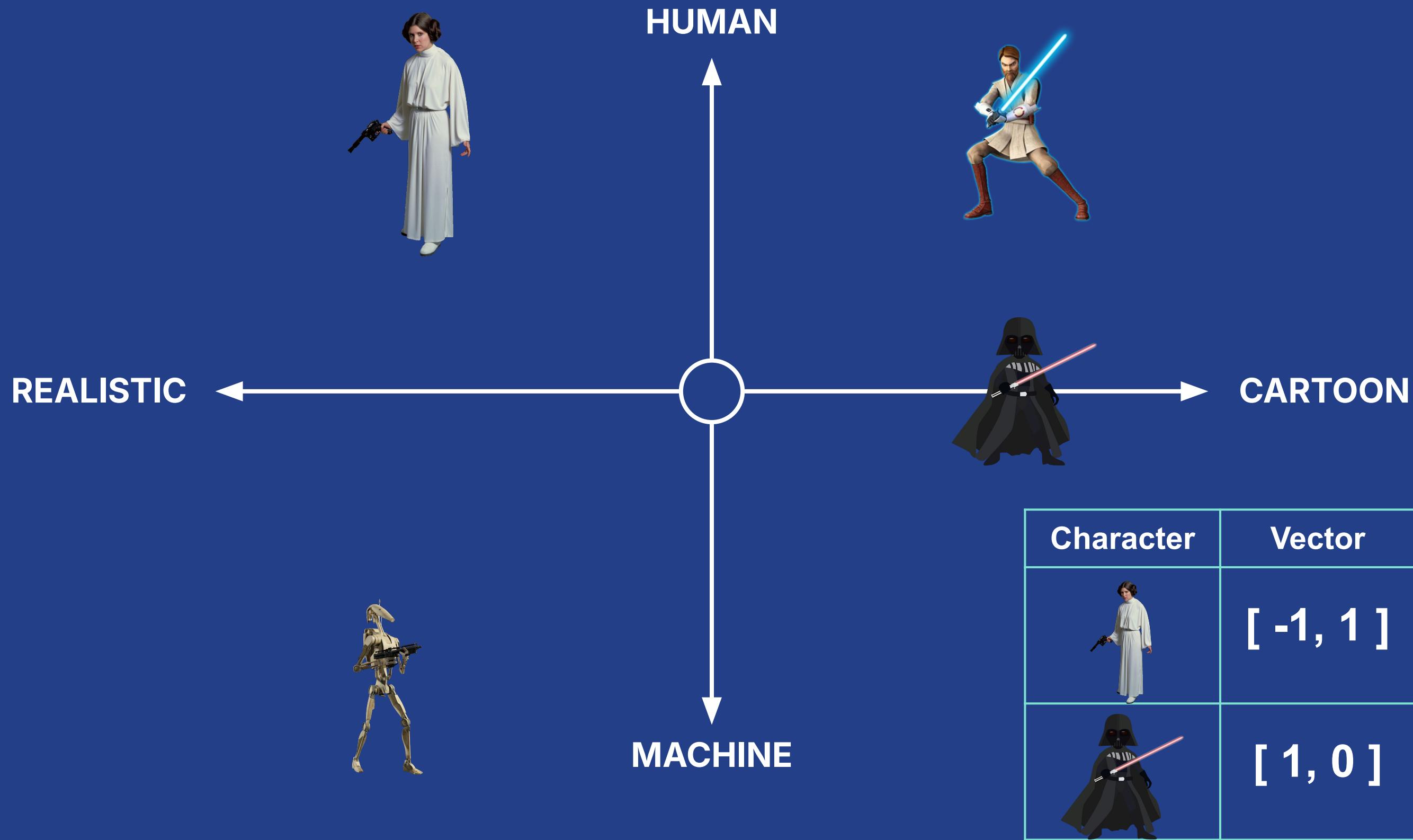
Embeddings represent your data

Example: 1-dimensional vector

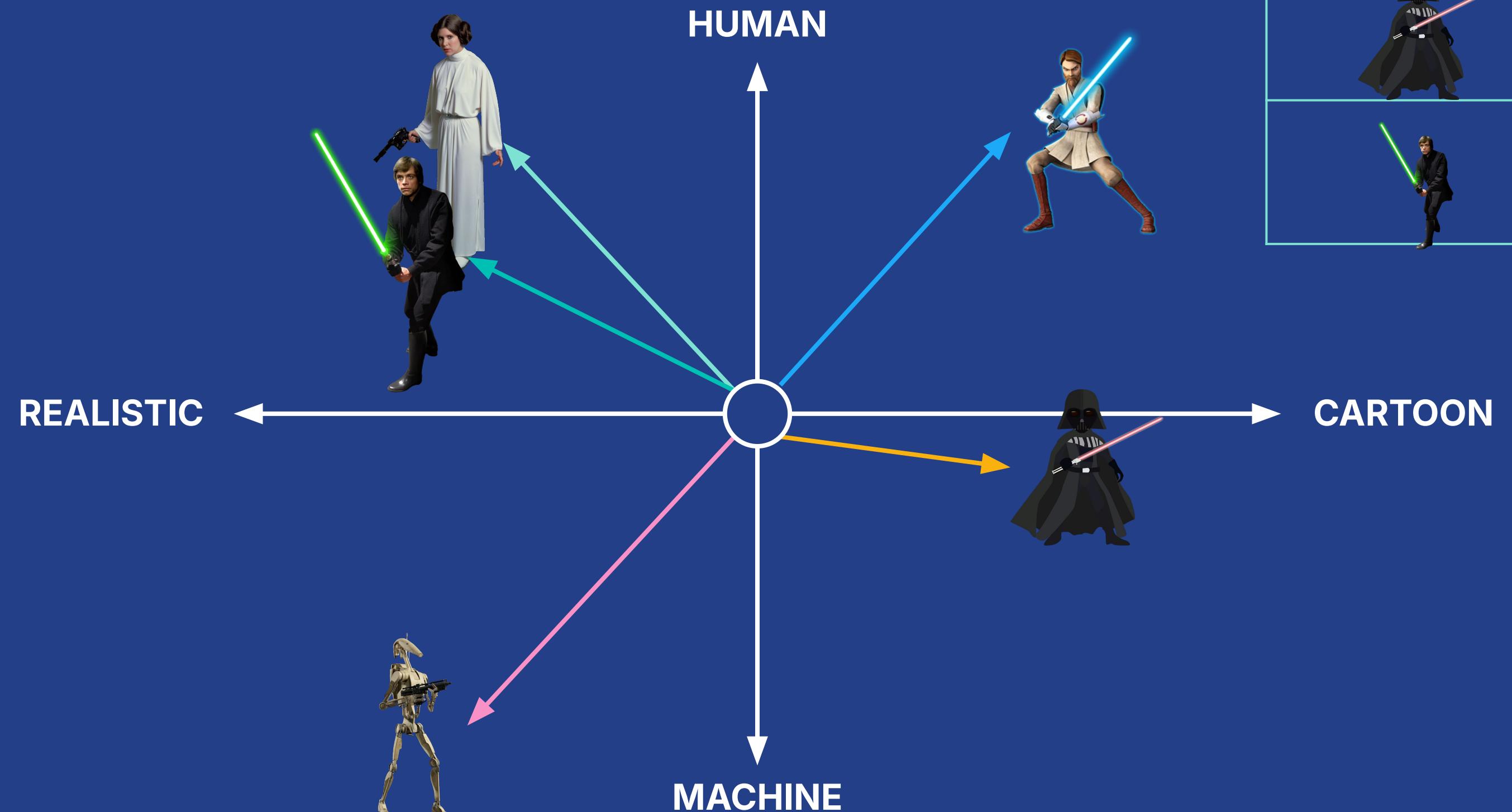


Character	Vector
	[-1]
	[1]

Multiple dimensions represent different data aspects

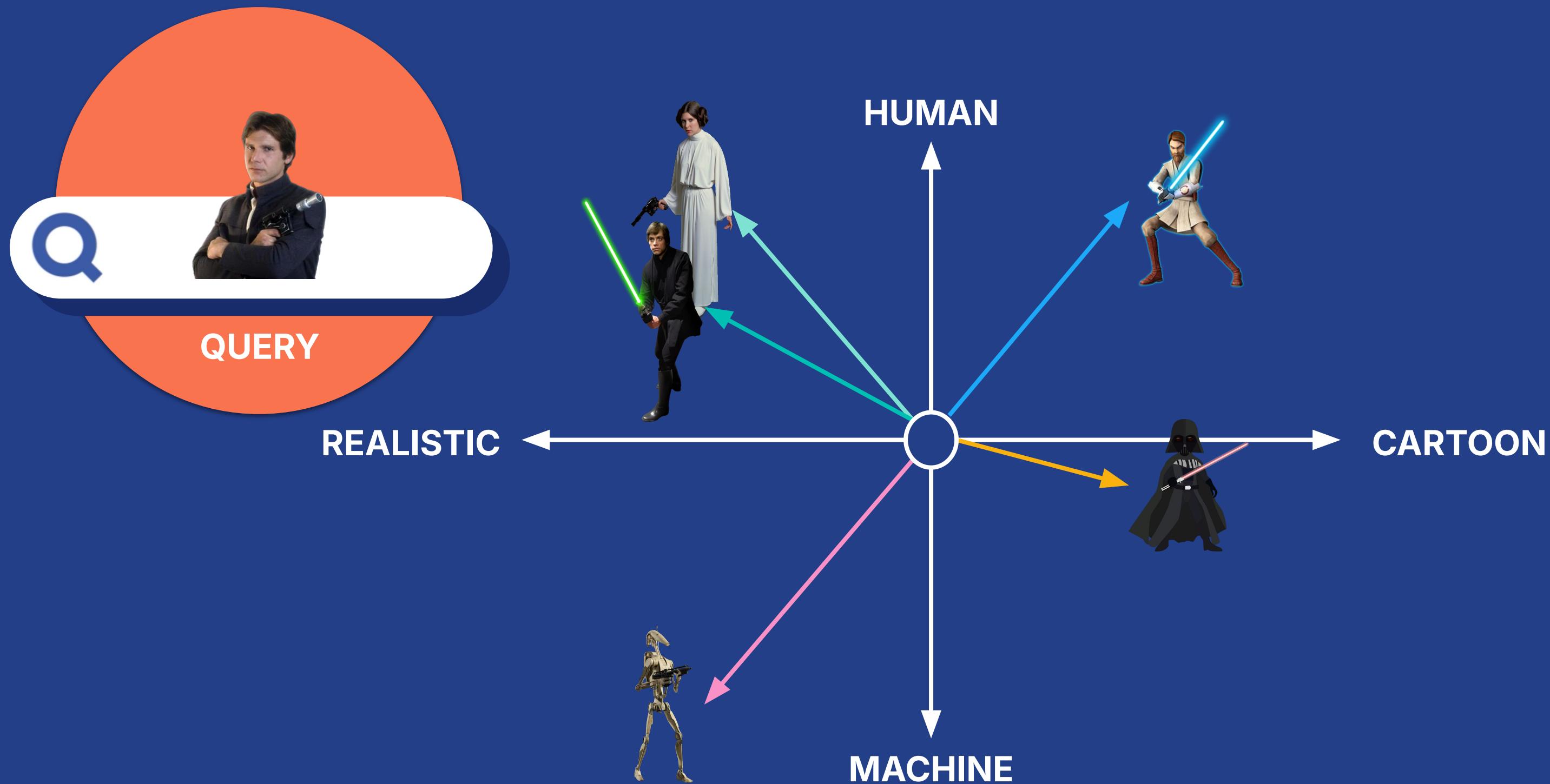


Similar data is grouped together



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

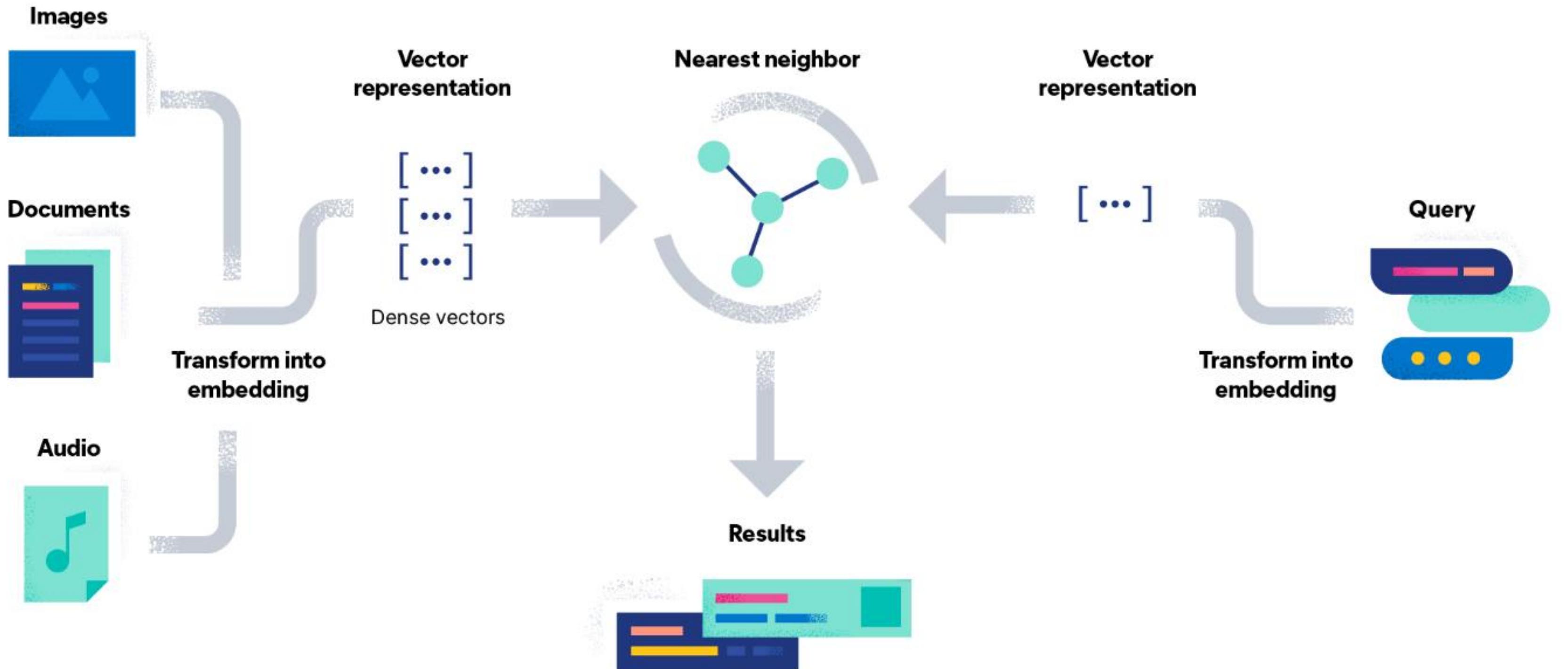
Vector search ranks objects by similarity (relevance) to the query



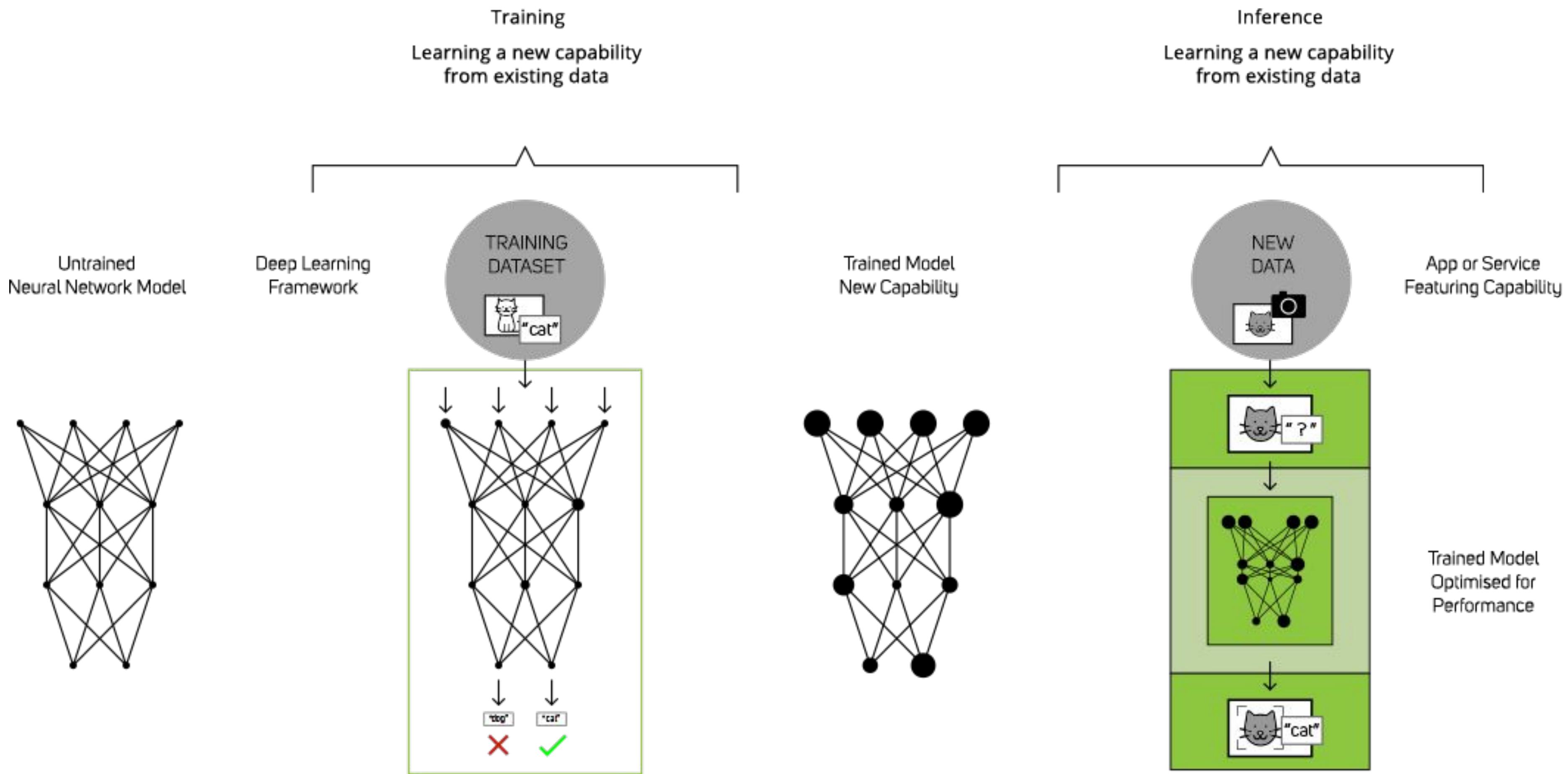
Relevance	Result
Query	
1	
2	
3	
4	
5	

Vector search conceptual architecture

Use vector nearest neighbor to generate a search ranking

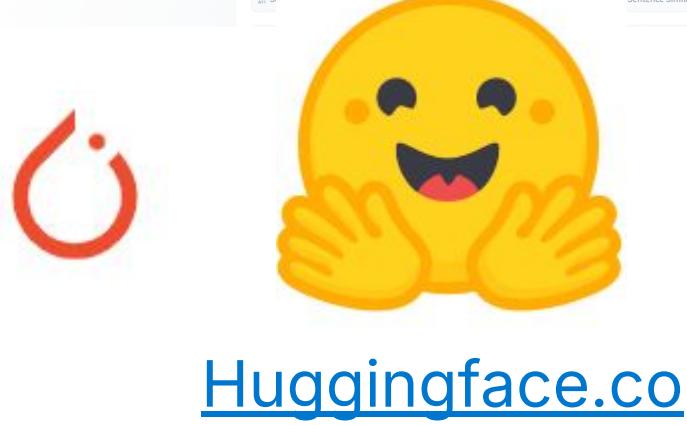


What is a Model?



Eland Imports PyTorch Models

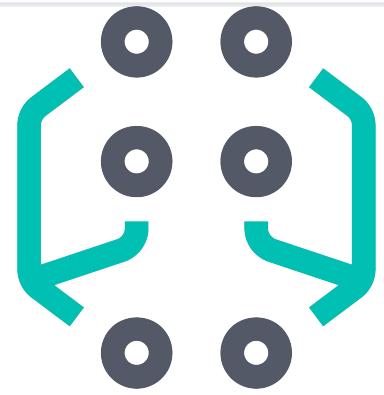
The screenshot shows the Hugging Face website's search interface. The search bar at the top contains the query "PyTorch". Below the search bar, there are several sections: "Tasks" (Fill-Mask, Question Answering, Summarization, Text Generation, Text2Text Generation, Token Classification, Translation, Zero-Shot Classification, Sentence Similarity), "Libraries" (PyTorch, TensorFlow, JAX), "Datasets" (wikipedia, common_voice, bookcorpus, glue, squad, deep_europarl_jrc_acquis, conll2003, oscar), "Languages" (en, es, fr, de, zh, sv, fi, ja), "Licenses" (apache-2.0, mit, cc-by-4.0), and "Other" (AutoNLP Compatible, Infinity Compatible). The main area displays a list of PyTorch models, such as gpt2, bert-base-uncased, distilbert-base-uncased, roberta-base, t5-base, Helsinki-NLP/opus-mt-zh-en, bert-base-chinese, sentence-transformers/multiqa-MiniLM-L6-cos-v1, bert-base-multilingual-cased, distilbert-base-uncased-finetuned-sst-2-english, xlm-roberta-base, distilbert-base-cased, bert-base-cased, roberta-large, and sentence-transformers/all-MiniLM-L6-v2.



```
$ eland_import_hub_model  
--url https://Cluster_URL  
--hub-model-id bert_model  
--task-type text_embedding  
--start
```



The screenshot shows the Elastic Model Management interface. The top navigation bar includes "Machine Learning", "Trained Models", and "Model Management". The "Model Management" tab is selected. Below it, the "Trained Models" section is visible, showing a table of models. The table columns include ID, Description, Type, State, and Created at. The table lists several models, such as "distilbert-base-uncased-finetuned-sst-2-english", "dsiml_bert-base-ner", "elastic_distilbert-base-cased-finetuned-comt03-english", "lang_ident_model_1", "sentence-transformers_clip-vit-b-32-multilingual-v1", "sentence_transformers_msmarco-MiniLM-L-12-v3", and "typeform_distilbert-base-uncased-mnli".

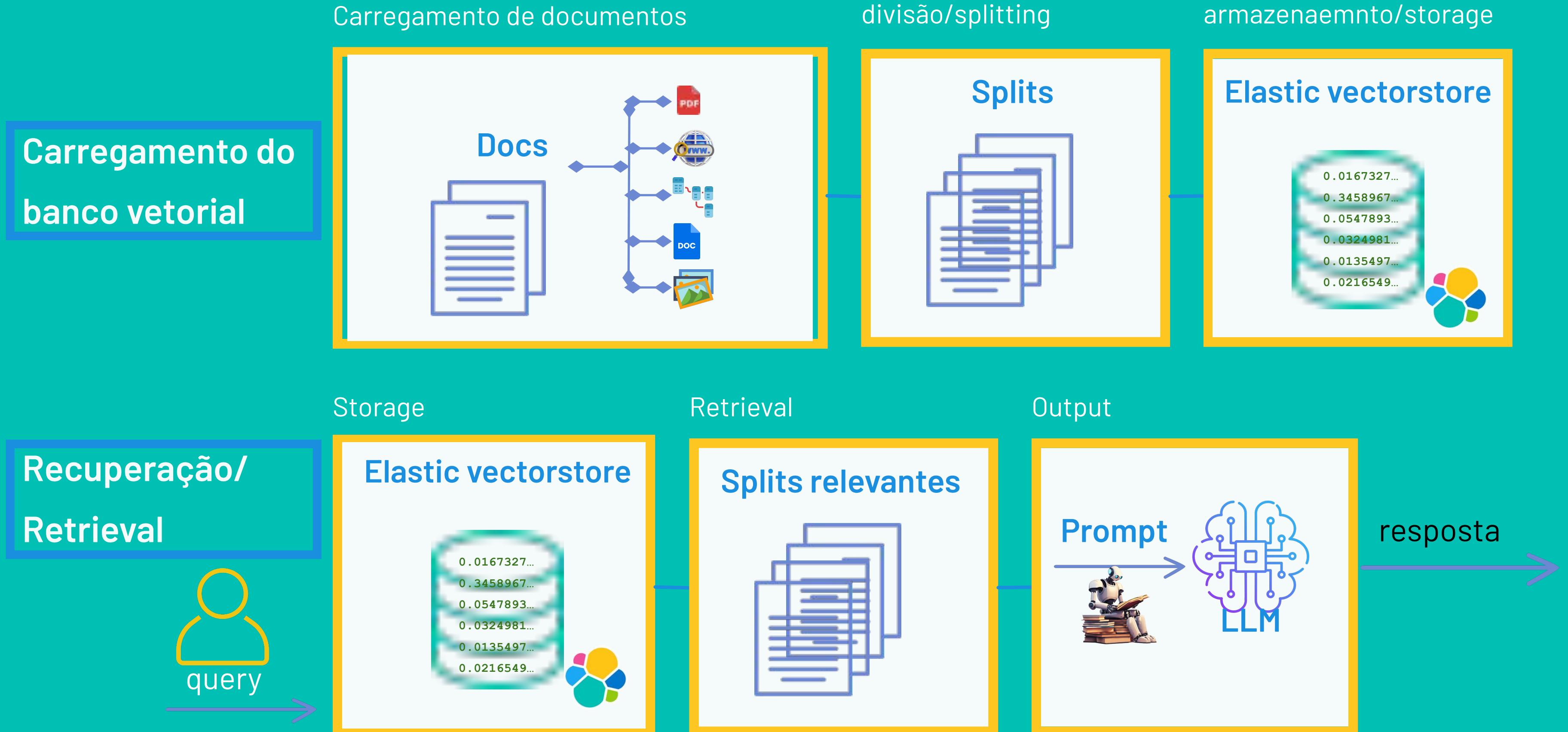


Inference, not training

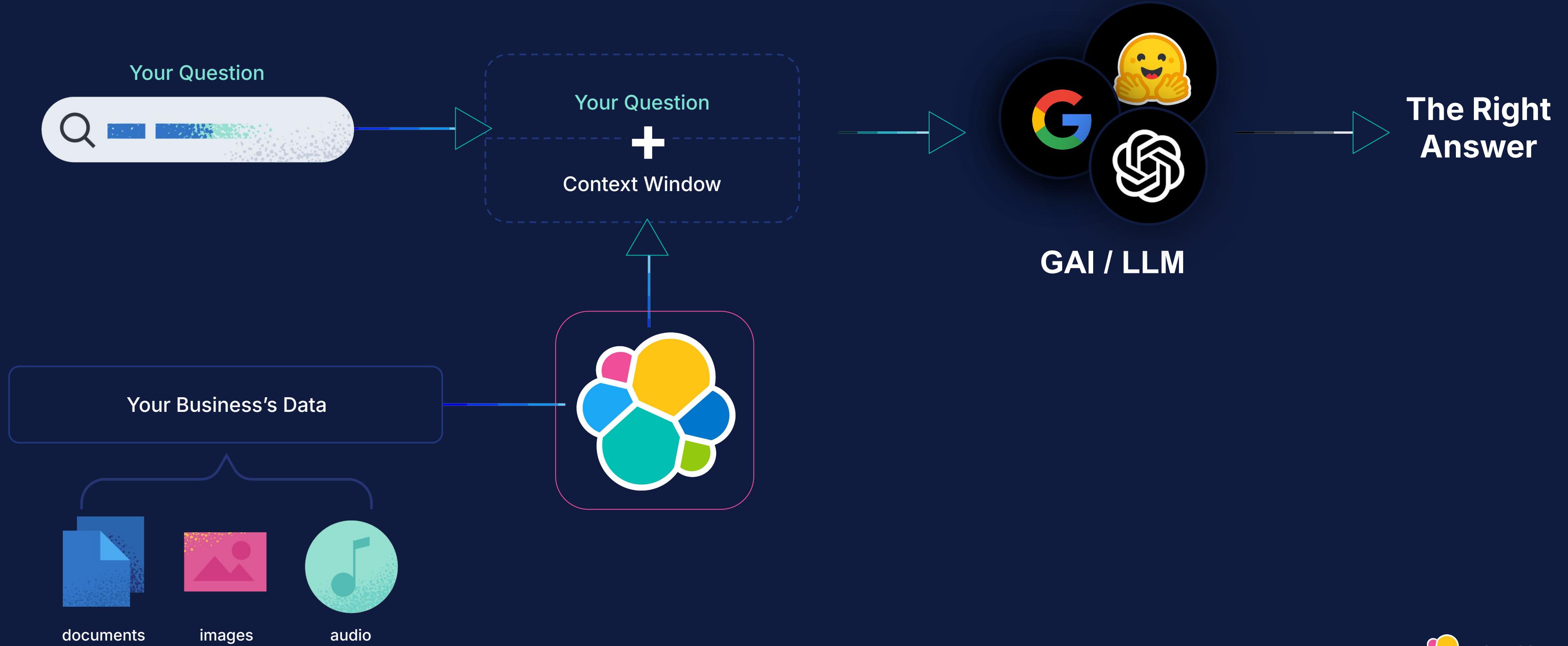


Chatbot Pipeline

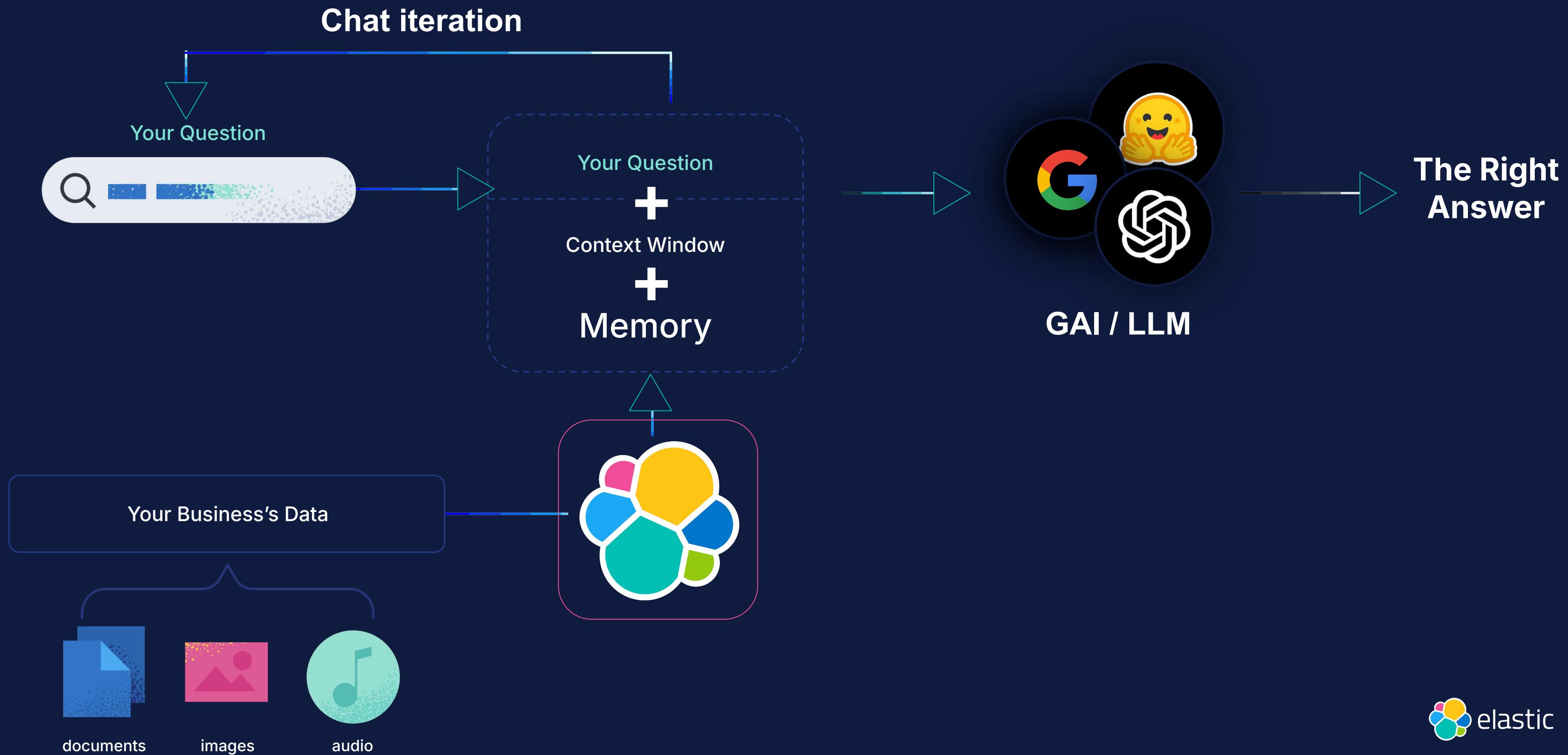
Converse com seus dados com este pipeline



RAG



Chatbot→RAG

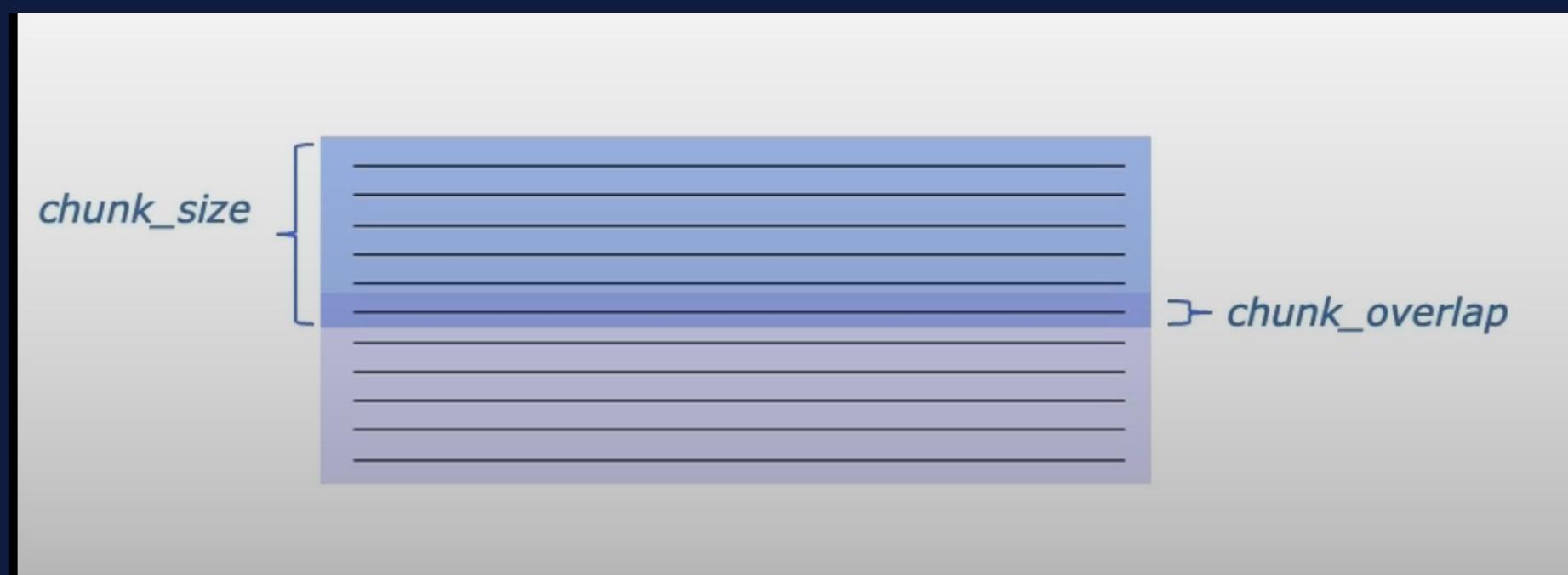


Exemplo de Splitter

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

Methods:

create_documents() - Create documents from a list of texts.
split_documents() - Split documents.



Tipos de Splitters

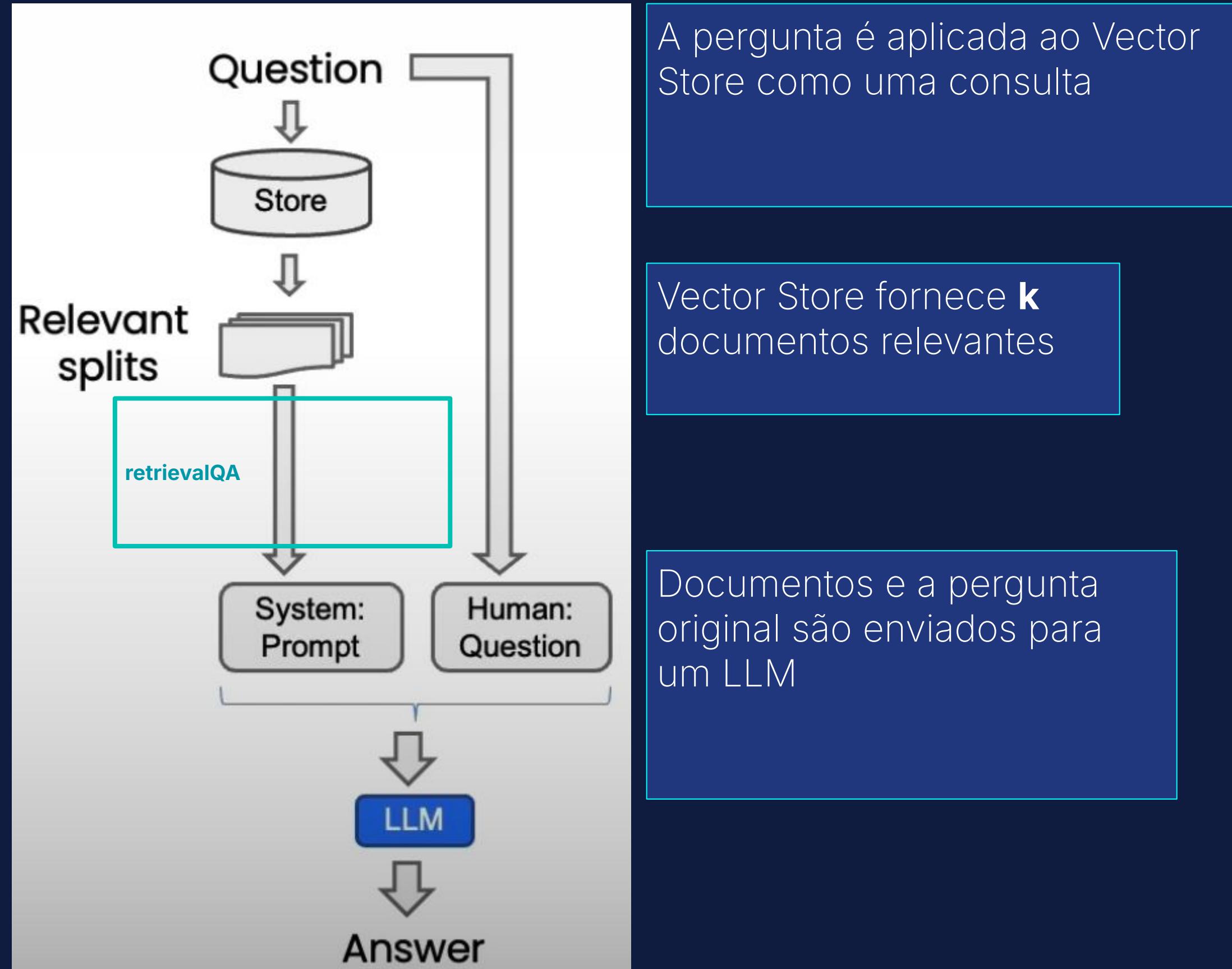
Langchain.text_splitter.

- **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

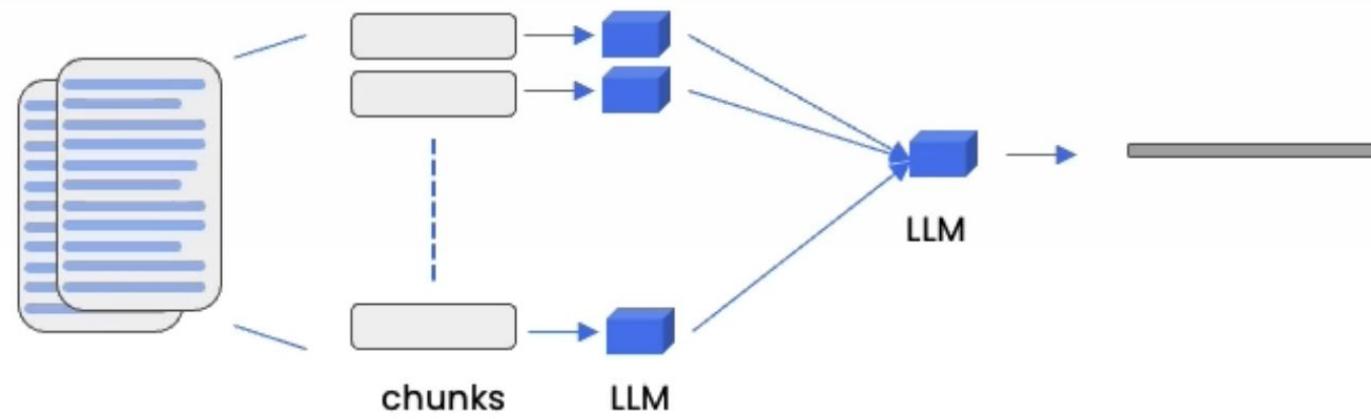
RetrievalQA chain

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

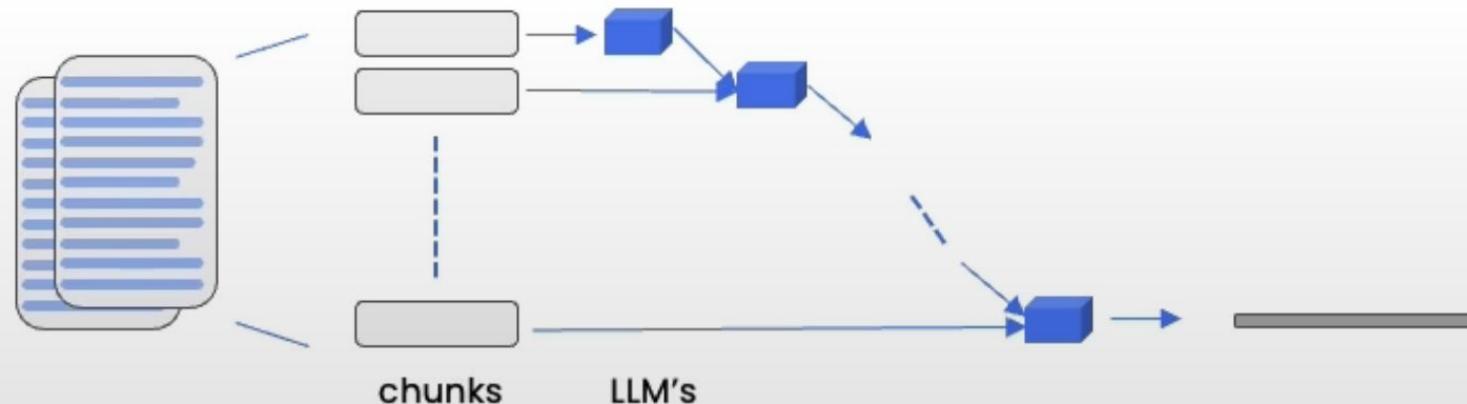


Tipos de Question Answerings

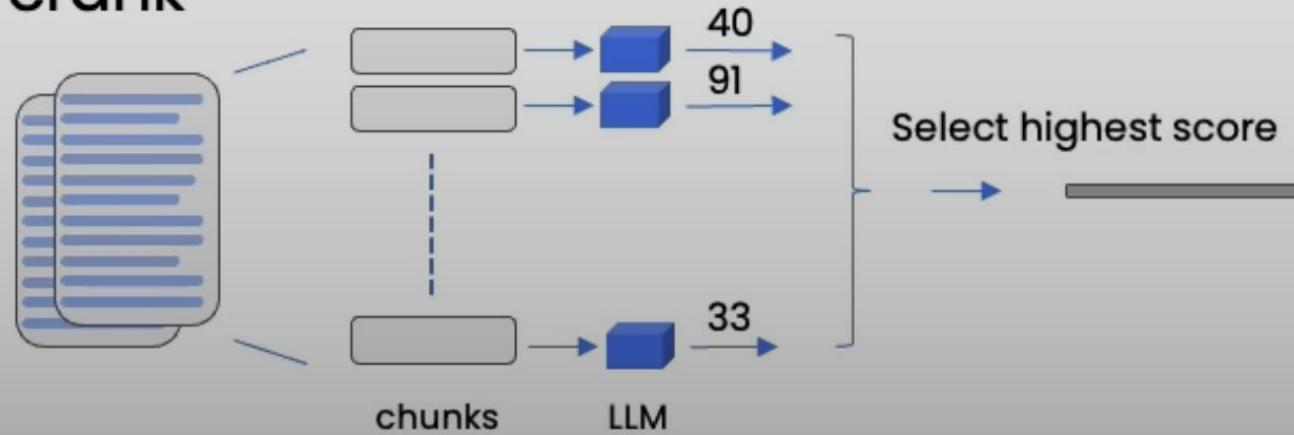
1. Map_reduce



2. Refine



3. Map_rerank



LangSmith Debugger

The screenshot shows the LangSmith Debugger interface for a 'RetrievalQA' run. The top navigation bar includes links for Home, Debug (which is selected), Testing, Monitoring, Datasets, and Documentation. The breadcrumb navigation shows the path: Debug > Project: default > Run: RetrievalQA.

Run: RetrievalQA

Chain Input: query: Is probability a class topic?

Chain Output: result: There is no clear answer to this question based on the given portion of the document. The document mentions familiarity with basic probability and statistics as a prerequisite for the class, and there is a brief mention of probability in the text, but it is not clear if it is a main topic of the class. The instructor mentions using a probabilistic interpretation to derive a learning algorithm, but does not go into further detail about probability as a topic.

Run Stats:

- RBC 1,503 → 87 1,590 tokens
- Start: 06/25/2023, 10:23:49 PM
- End: 06/25/2023, 10:24:01 PM
- Response time: 11.77s

Child Runs:

- RetrievalQA (11.77s) RBC 1,503 → 87
 - MapReduceDocumentsChain (11.63s) RBC
 - LLMChain (8.26s) RBC 1,267 → 0
 - ChatOpenAI (RBC 376 → 0)
 - ChatOpenAI (RBC 372 → 0)
 - ChatOpenAI (RBC 376 → 0)
 - ChatOpenAI (RBC 143 → 0)
 - StuffDocumentsChain (3.36s) RBC
 - LLMChain (3.35s) RBC 236 → 8
 - ChatOpenAI (RBC 2 → 0)

Run Metadata:

- completion_tokens: 87
- prompt_tokens: 1503
- total_tokens: 1590

RUNTIME:

- library: "langchain"
- library_version: "0.0.213"

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"

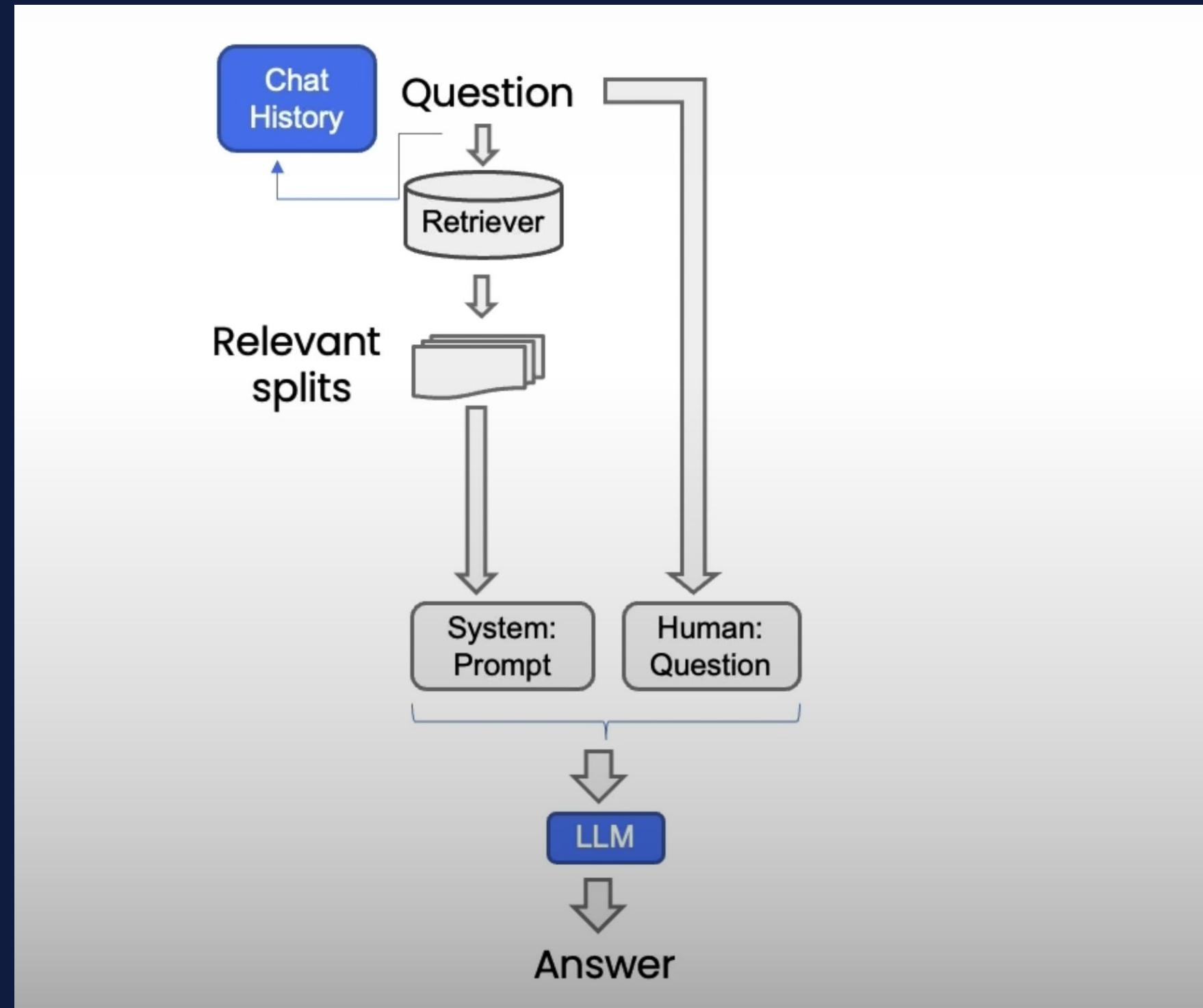
Histórico de Conversa

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

Desculpe, não faço ideia do que você está falando!

ConversationRetrievalChain

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



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>

Recursos para desenvolvedores: Elasticsearch Labs

elastic.co/search-labs

BLOG / ML RESEARCH

Evaluating RAG: A journey through metrics



In 2020, Meta published a paper titled "[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)". This paper introduced a method for expanding the knowledge of Language

github.com/elastic/elasticsearch-labs

 **elasticsearch-labs** Public

About
Elasticsearch Guides, Notebooks & Example Apps for Search Applications

search-labs.elastic.co/search-labs

python search elasticsearch ai
vector applications openai elastic
chatlog chatgpt langchain
openai-chatgpt genai genaistack
vectordatabase

Readme Apache-2.0 license
Security policy
Activity
109 stars
178 watching
40 forks
Report repository

Languages

Jupyter Notebook	93.7%
Python	2.9%
TypeScript	1.3%
Handlebars	0.1%
JavaScript	1.6%
CSS	0.2%
Other	0.2%

Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene



Recursos para desenvolvedores: Junte-se à Comunidade Elastic

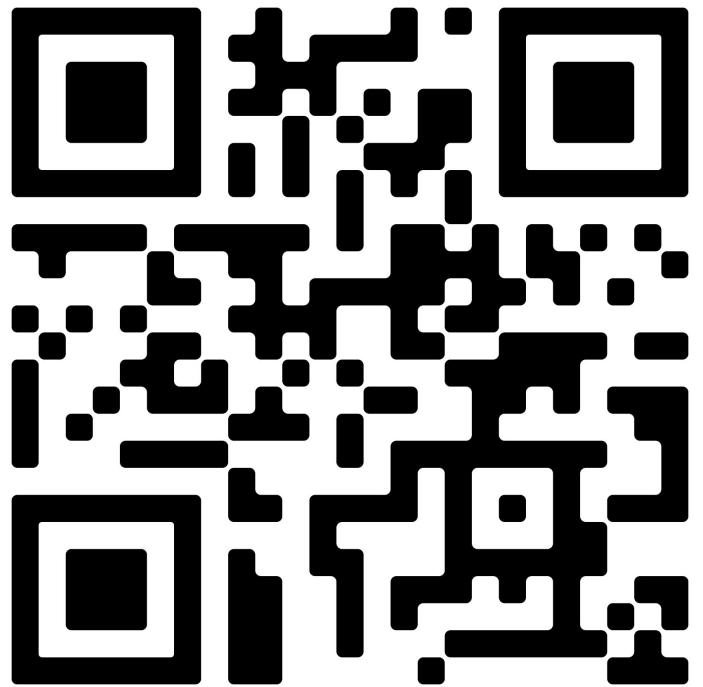
Elastic User Groups

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



Compartilhe o seu conhecimento

Fóruns de discussão

Encontre conselhos e ajude o próximo.

[Faça uma pergunta →](#)

Slack

Entre no nosso Elastic Slack, que está crescendo rapidamente, para conversar com outros usuários e pedir conselhos.

[Participe da conversa →](#)

Colaboradores

Seja reconhecido(a) pelas suas contribuições para a comunidade Elastic. Queremos ver o seu nome no placar dos líderes!

[Comece a contribuir →](#)



Obrigado

Alex Salgado

Senior Developer Advocate, Elastic



@alexsalgadopro f



salgado



@alexsalgadoprof



/in/alex-salgado/

