DevFest

DevFest Cerrado

Vini Caridá

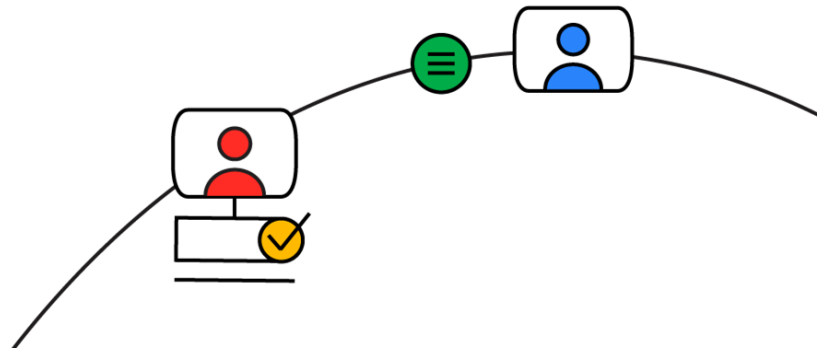
Usando Tecnologia, Dados e AI para alavancar negócios e impactar pessoas para um mundo mais justo e evoluido.

Google Developer Groups

DevFest 2024 Google Developer Groups Cerrado

VINICARIDÁ Human-Centered Data & AI

Desvendando a GenAI: Da Teoria à Aplicações com Gemini e Gemma



Experts Google Developers

VINICARIDÁ
Human-Centered Data & AI

Vinicius Caridá, Ph.D.

- Executive Specialist, Artificial Intelligence and Data - Itaú
- MBA Professor – FIAP and ESPM

@vinicius caridá @vfcarida @vinicius caridá @vfcarida @vinicius caridá @vfcarida

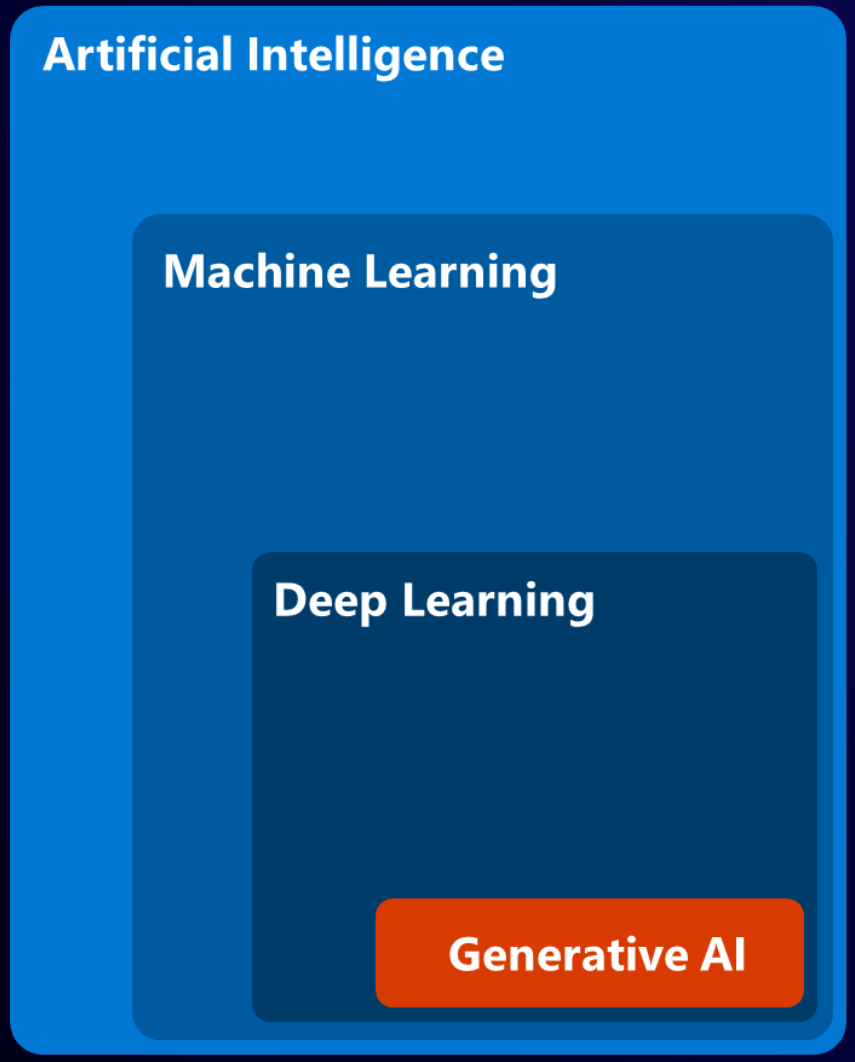
VINICARIDÁ
Human-Centered Data & AI

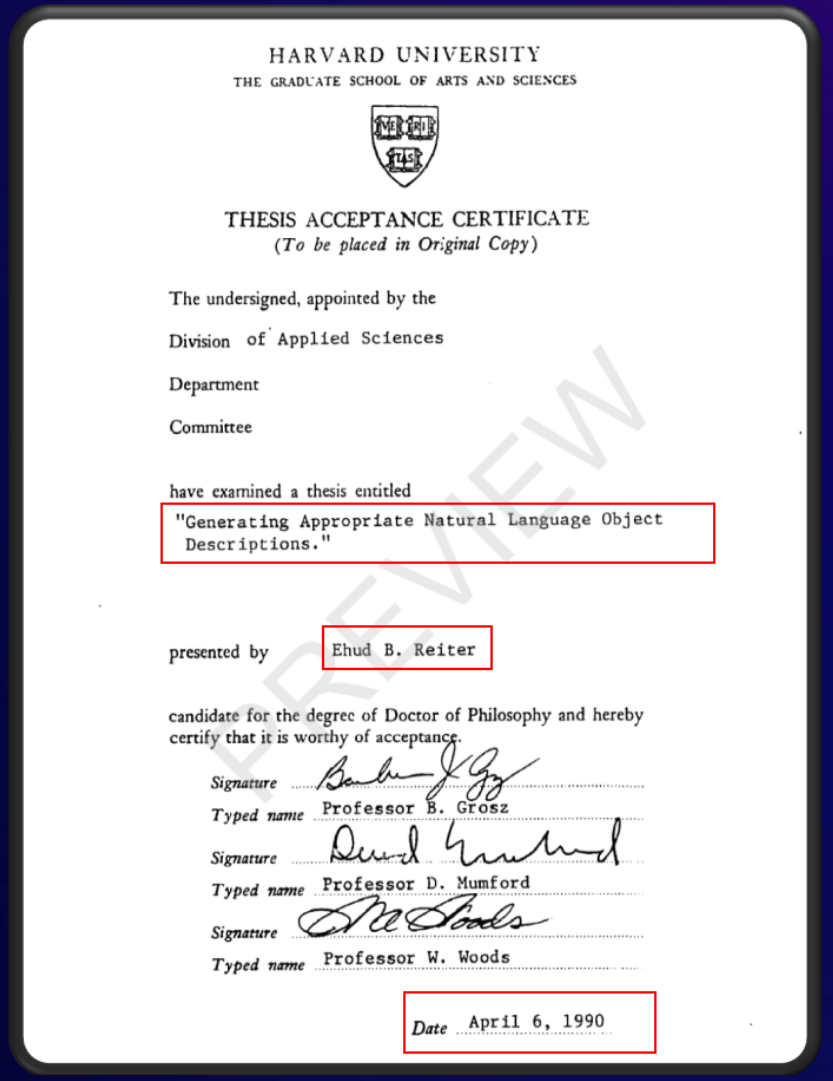
01 GenAI, Hype ou Usual?

DevFest 2024 Cerrado

Generative AI

VINICARIDÁ
Human-Centered Data & AI





HARVARD UNIVERSITY
THE GRADUATE SCHOOL OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE
(To be placed in Original Copy)

The undersigned, appointed by the

Division of Applied Sciences

Department

Committee

have examined a thesis entitled

"Generating Appropriate Natural Language Object Descriptions."

presented by Ehud B. Reiter

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature .....
Typed name Professor B. Grosz

Signature .....
Typed name Professor D. Mumford

Signature .....
Typed name Professor W. Woods

Date April 6, 1990

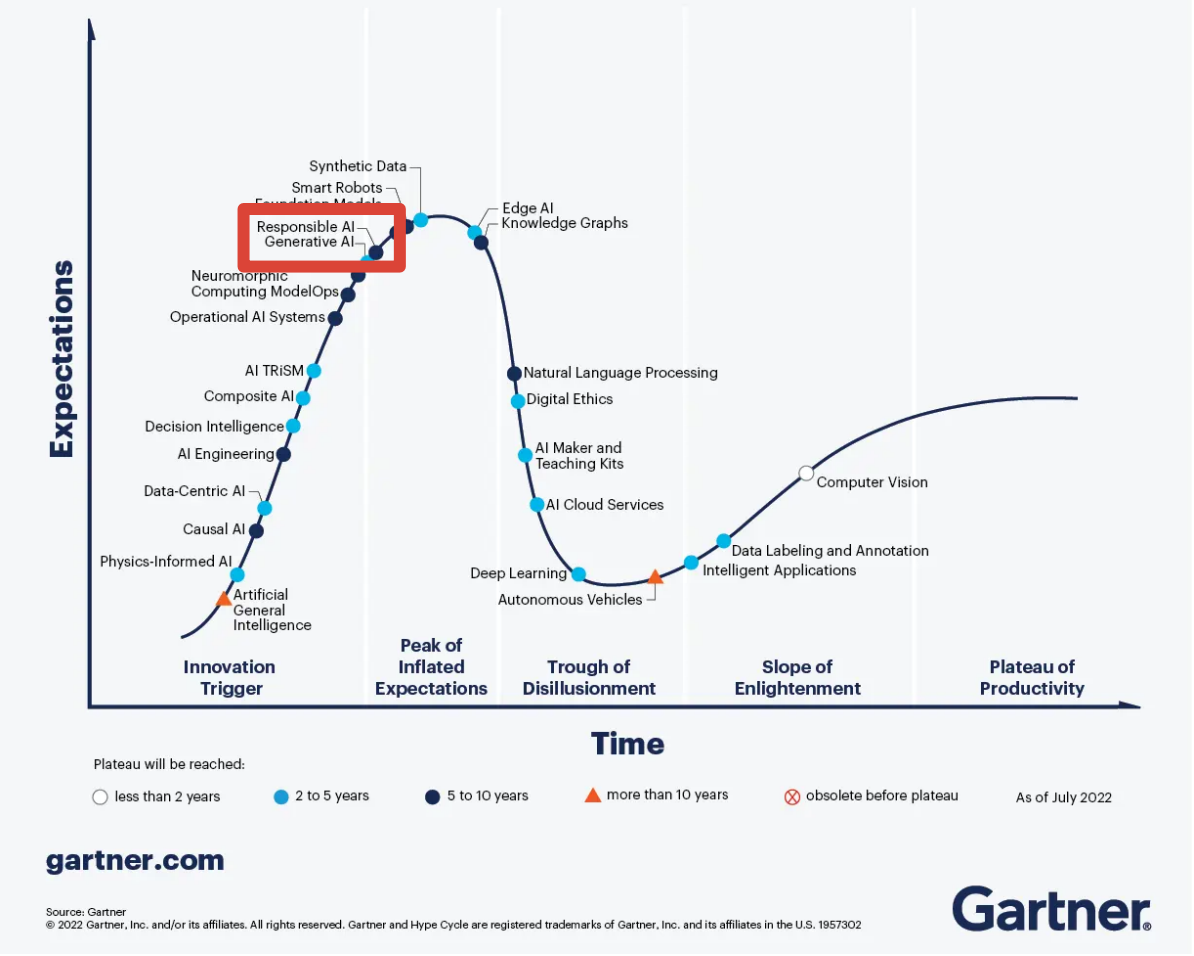
PREVIEW

https://www.proquest.com/openview/a17fb9188f537c2ab3df4091453547ba/

Generative AI | Hype cycle for AI

VINICARIDÁ
Human-Centered Data & AI

2022

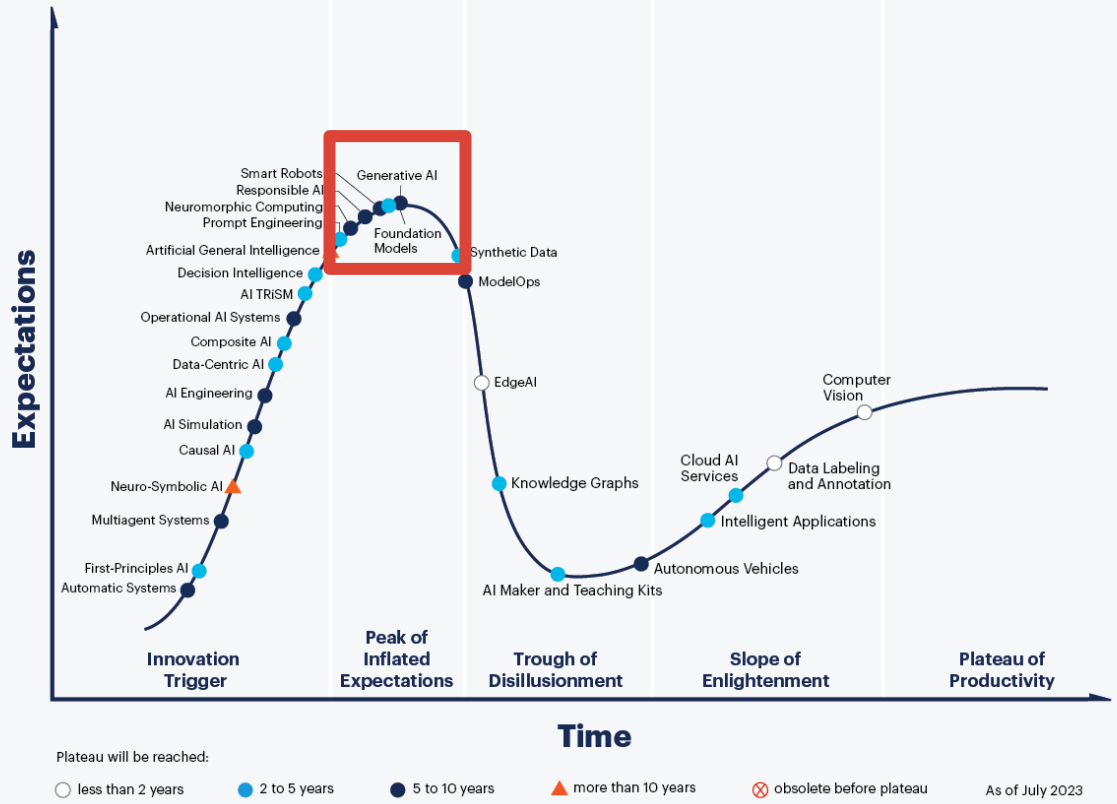


gartner.com

Source: Gartner
© 2022 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S. 1957302

Gartner

2023



gartner.com

Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2079794

Gartner

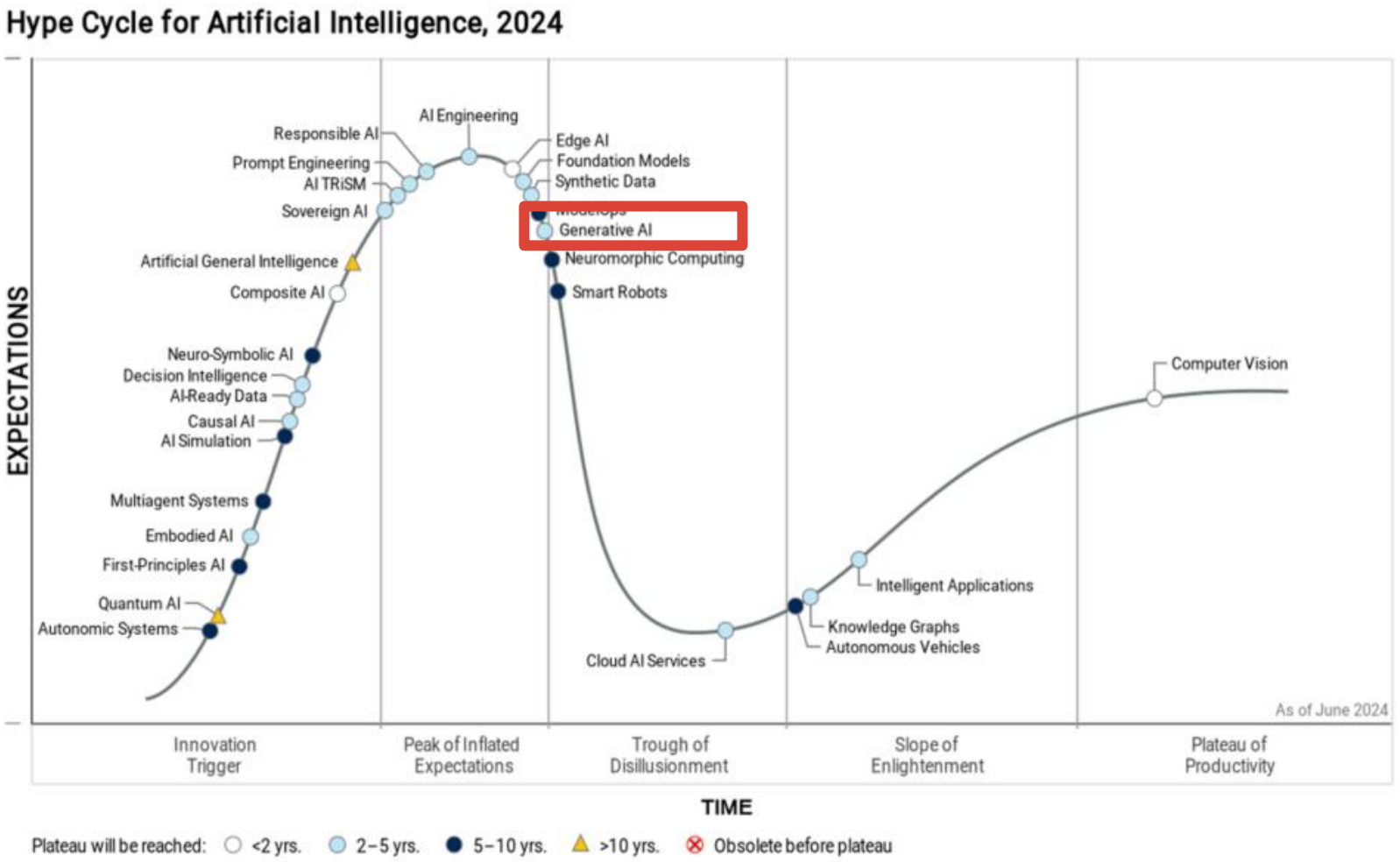
Experts
Google Developers

DevFest 2024

Generative AI | Hype cycle for AI

VINICARIDÁ
Human-Centered Data & AI

2024



Experts
Google Developers

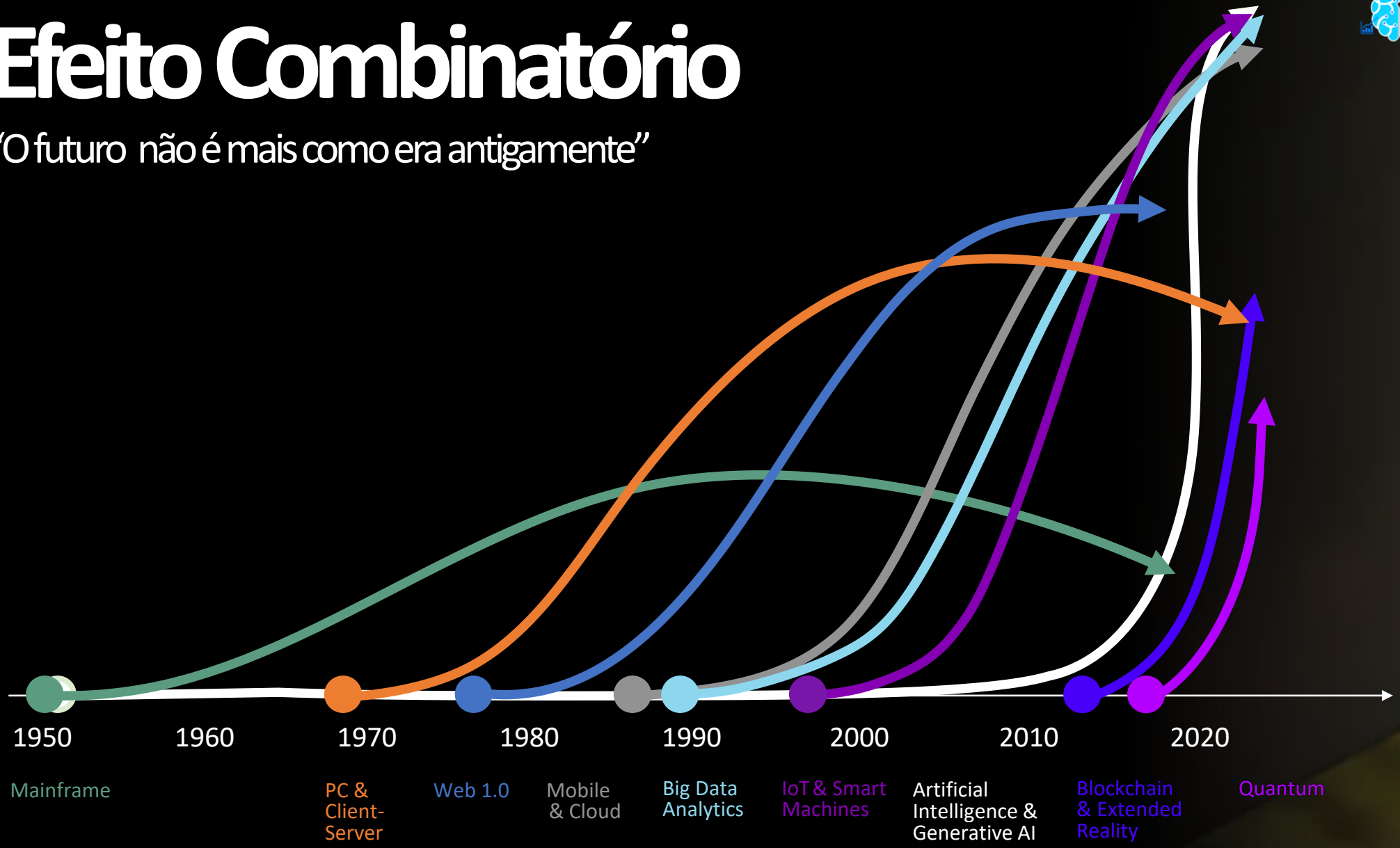
Gartner

DevFest 2024

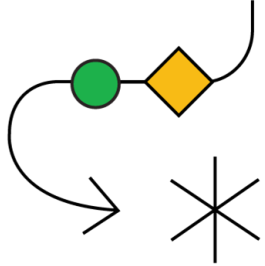
Tecnologias emergindo

Efeito Combinatório

"O futuro não é mais como era antigamente"

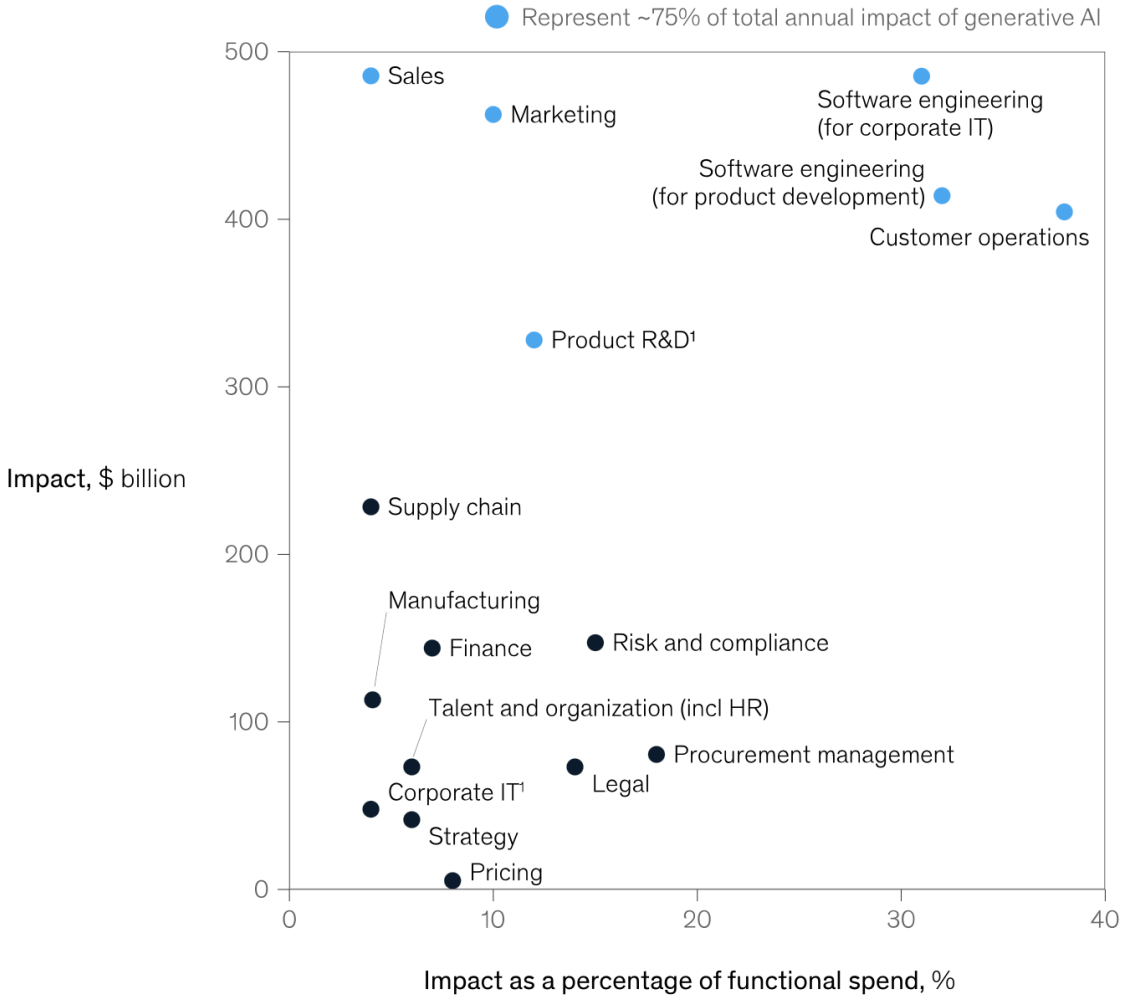


Fonte: Accenture Tech Vision 2022



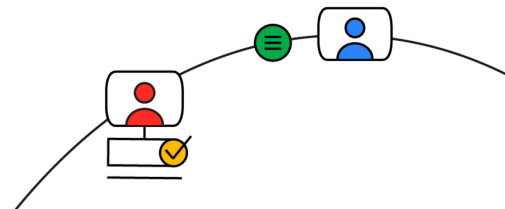
VINICARIDÁ
Human-Centered Data & AI

Using generative AI in just a few functions could drive most of the technology's impact across potential corporate use cases



https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/The-economic-potential-of-generative-AI-The-next-productivity-frontier

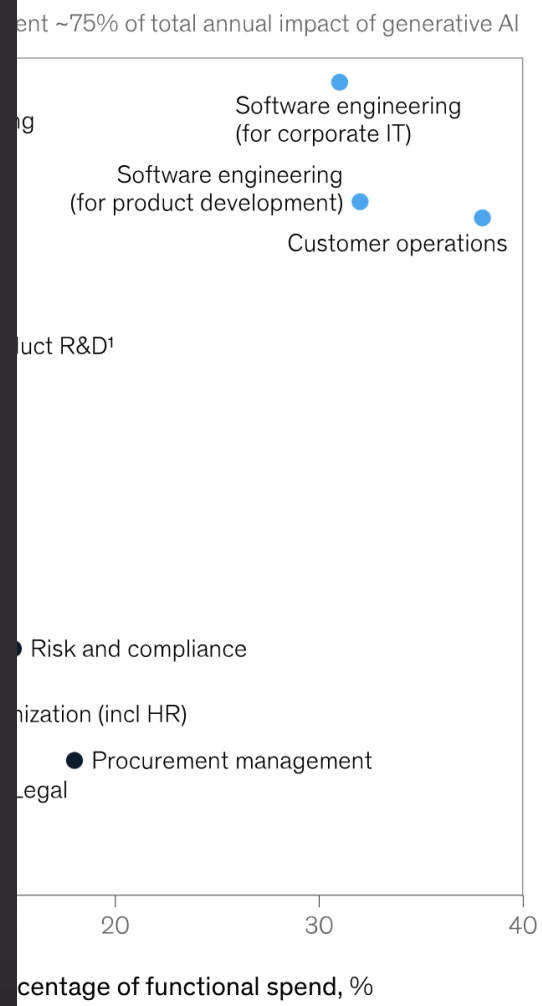
DevFest Cerrado 2024 Google Developer Groups



VINICARIDÁ AI

Using genera
just a few fur
could drive m
technology's
across poten
corporate us

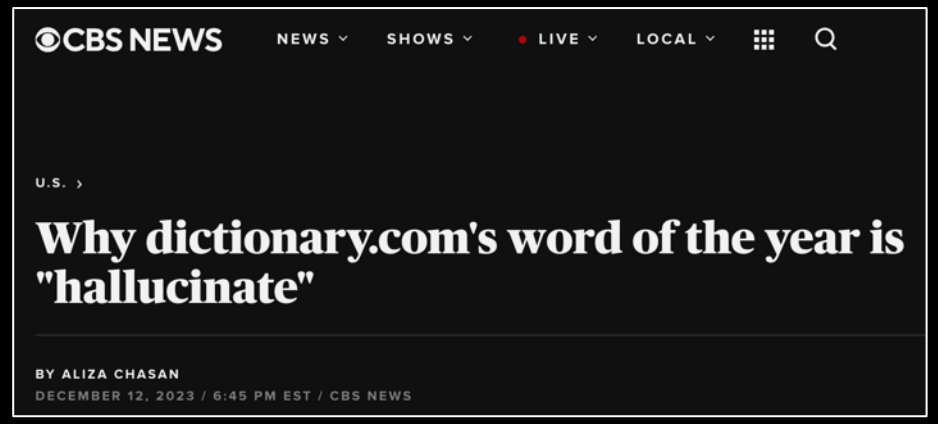
https://www.mckinsey.com/capabilities/mc
economic-potential-of-generative-AI-The-ne



DevFest 2024 Google Developer Groups Cerrado

VINI CARIDÁ
Human-Centered Data & AI

GenAI na nossa rotina


CBS NEWS NEWS SHOWS LIVE LOCAL

U.S. ›

Why dictionary.com's word of the year is "hallucinate"

BY ALIZA CHASAN
DECEMBER 12, 2023 / 6:45 PM EST / CBS NEWS

Um novo linguajar nos negócios. **&** Uma preocupação constante.

Sinais de ruptura

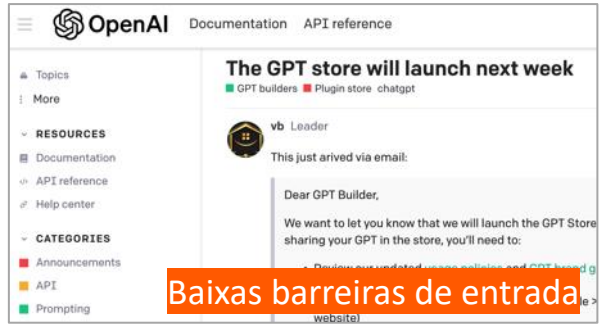

Big-name financial institutions are developing LLMs purpose-built for financial services

Bloomberg
Bloomberg is developing BloombergGPT – an LLM built for financial tasks like sentiment analysis and news classification.

JPMorgan Chase & Co.
JP Morgan Chase is developing IndexGPT, an AI chatbot for investment advice and selection.

INTUIT
Intuit partnered with OpenAI to develop GenOS – a genAI operating system that will power customer experiences across its product suite.

Investimentos do setor financeiro


OpenAI Documentation API reference

Topics
More
RESOURCES
Documentation
API reference
Help center
CATEGORIES
Announcements
API
Prompting

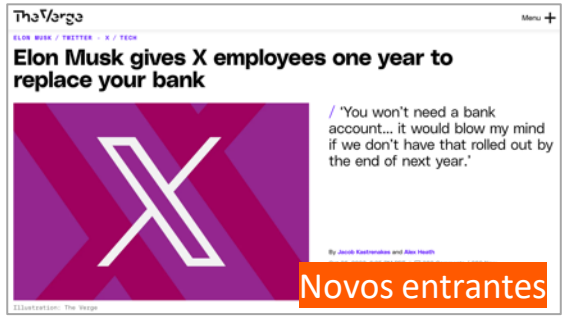
The GPT store will launch next week
GPT builders Plugin store chatgpt

vb Leader
This just arived via email:

Dear GPT Builder,

We want to let you know that we will launch the GPT Store sharing your GPT in the store, you'll need to:

Baixas barreiras de entrada


The Verge

Elon Musk gives X employees one year to replace your bank

'You won't need a bank account... it would blow my mind if we don't have that rolled out by the end of next year.'

Novos entrantes


THE FINANCIAL BRAND banking transforme

SUBSCRIBE FREE ALL ARTICLES FORUM 2024 WEBINARS PODCASTS

In 2024, Gen Z Will Get Its Money Advice from TikTok. Should Banks Jump into the Conversation?

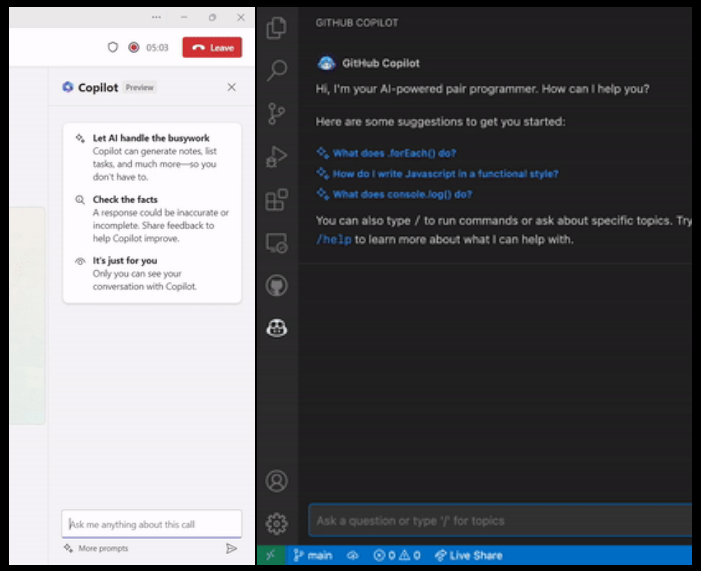
SUBSCRIBE NOW

Mudança de comportamento

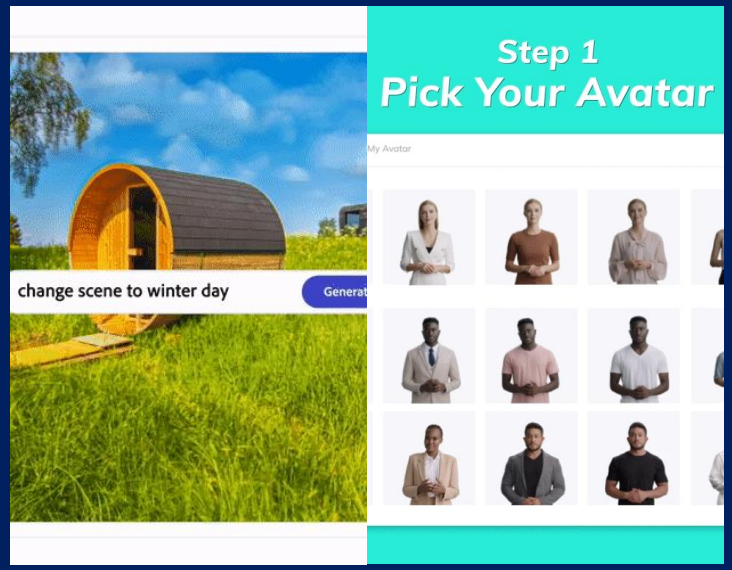
O que é multimodal?

VINI CARIDÁ
Human-Centered Data & AI

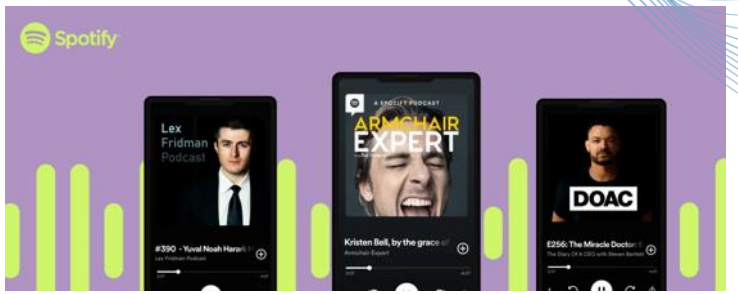
Textos & Código

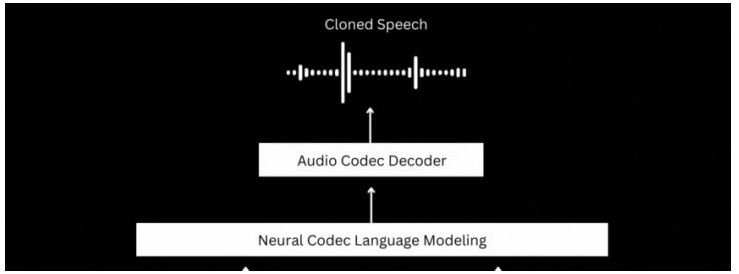


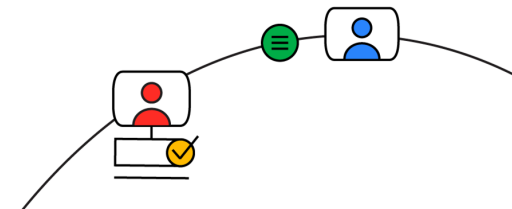
Imagens, Vídeos



Áudio



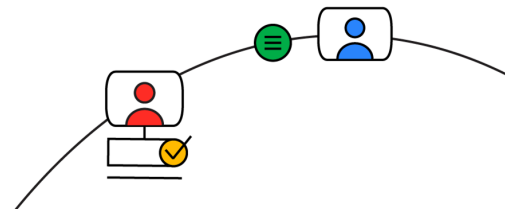




DevFest Cerrado 2024 Google Developer Groups

VINICARIDÁ
Human-Centered Data & AI

Experts
Google Developers



DevFest Cerrado 2024 Google Developer Groups

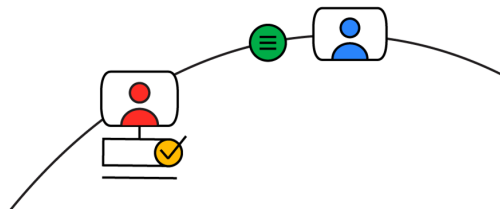
VINICARIDÁ Human-Centered Data & AI

Experts Google Developers

DevFest Cerrado 2024 Google Developer Groups

VINICARIDÁ
Human-Centered Data & AI

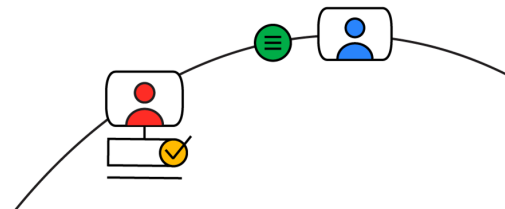
Experts
Google Developers

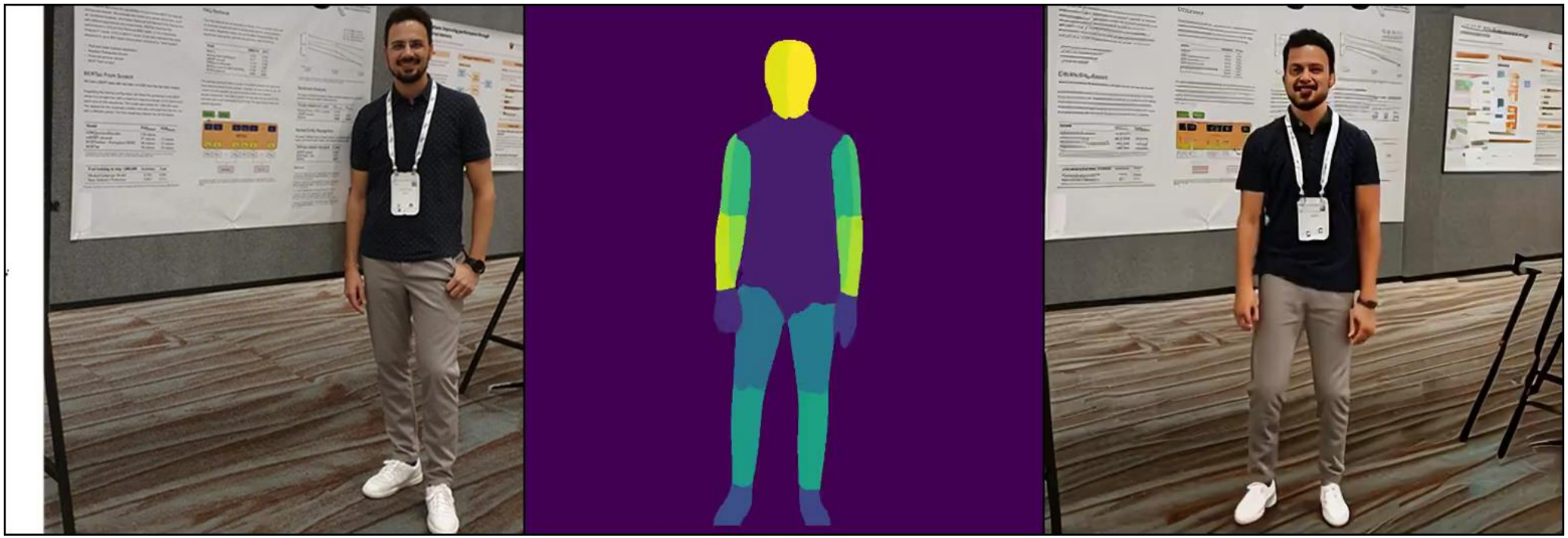


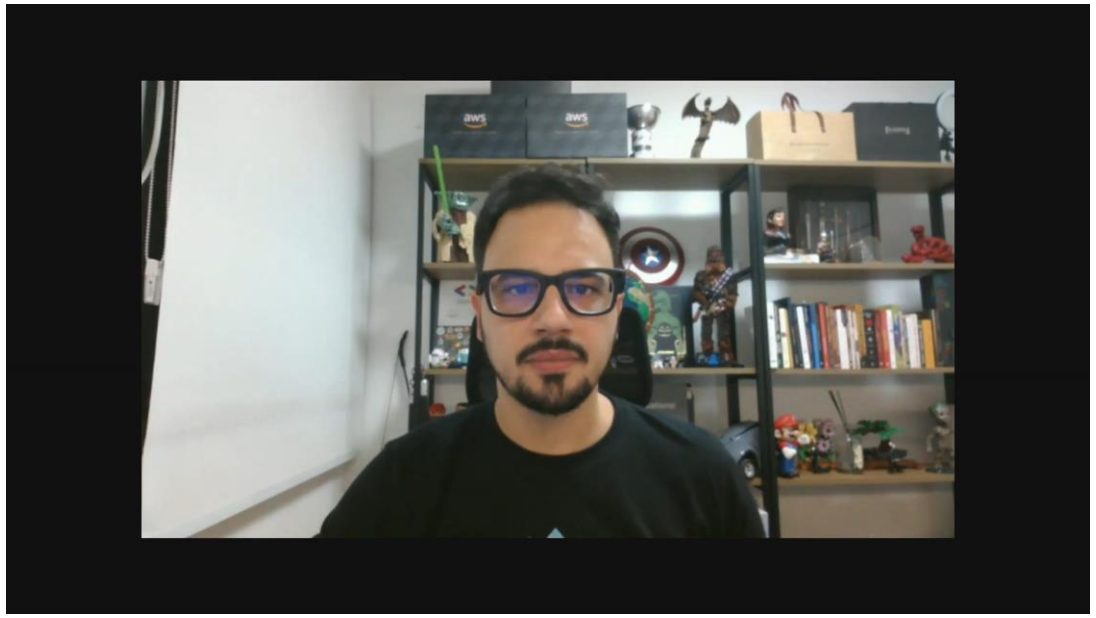
DevFest 2024 Google Developer Groups Cerrado

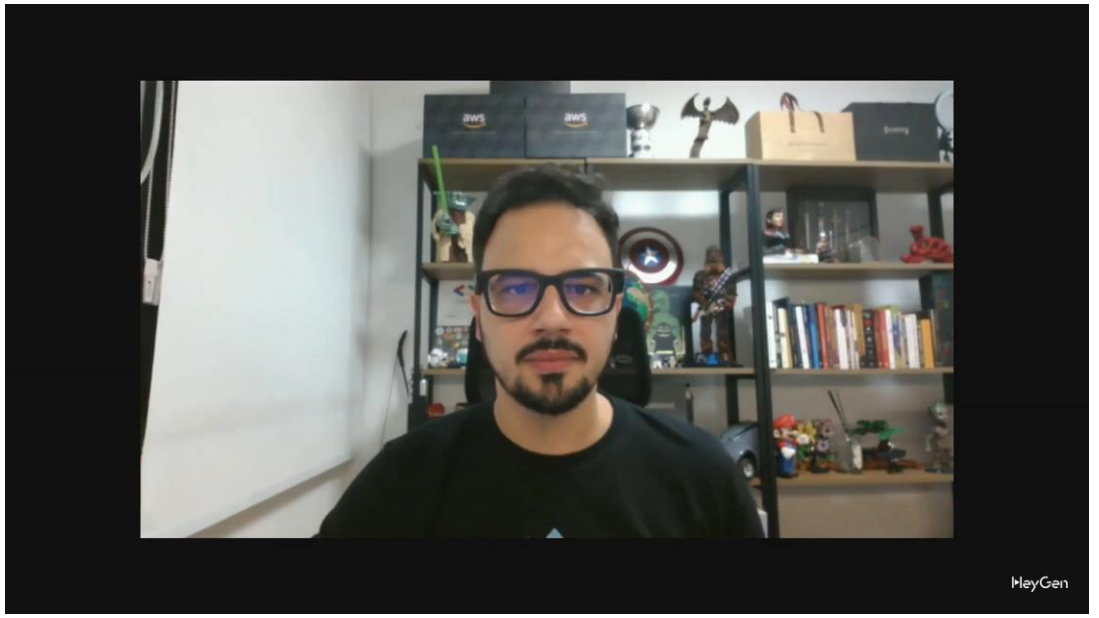
VINICARIDÁ
Human-Centered Data & AI

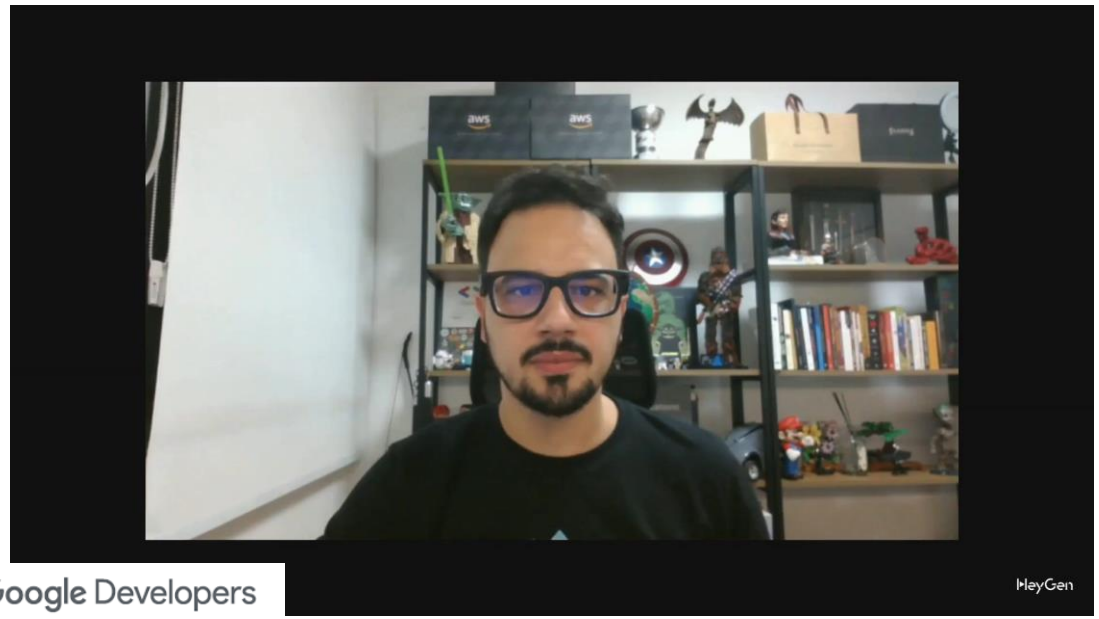
Experts
Google Developers

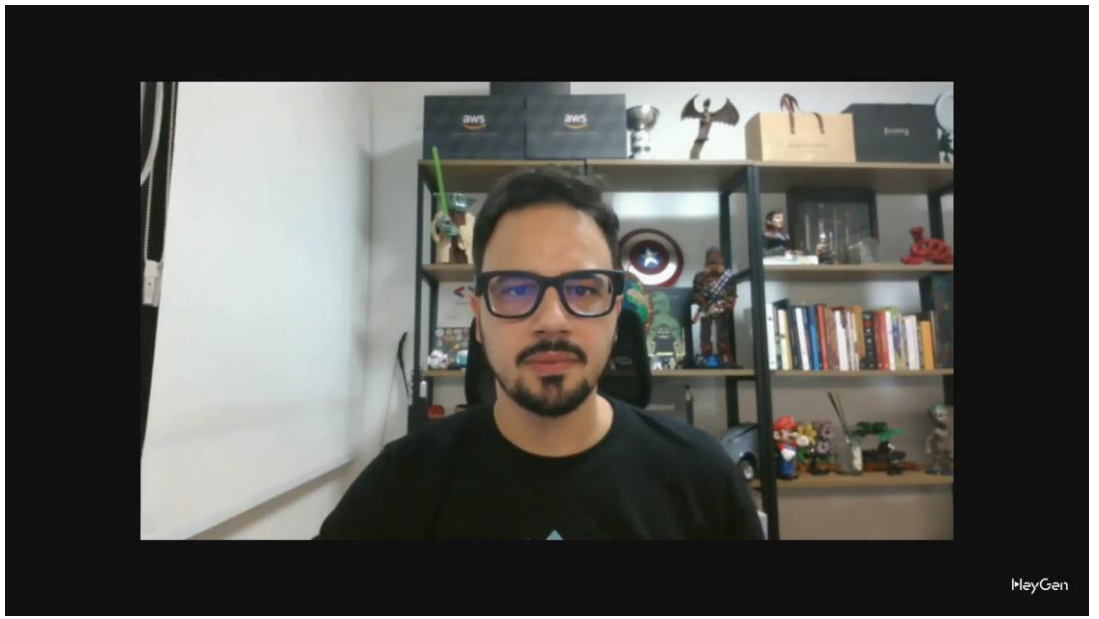







HeyGen


HeyGen


HeyGen


runway

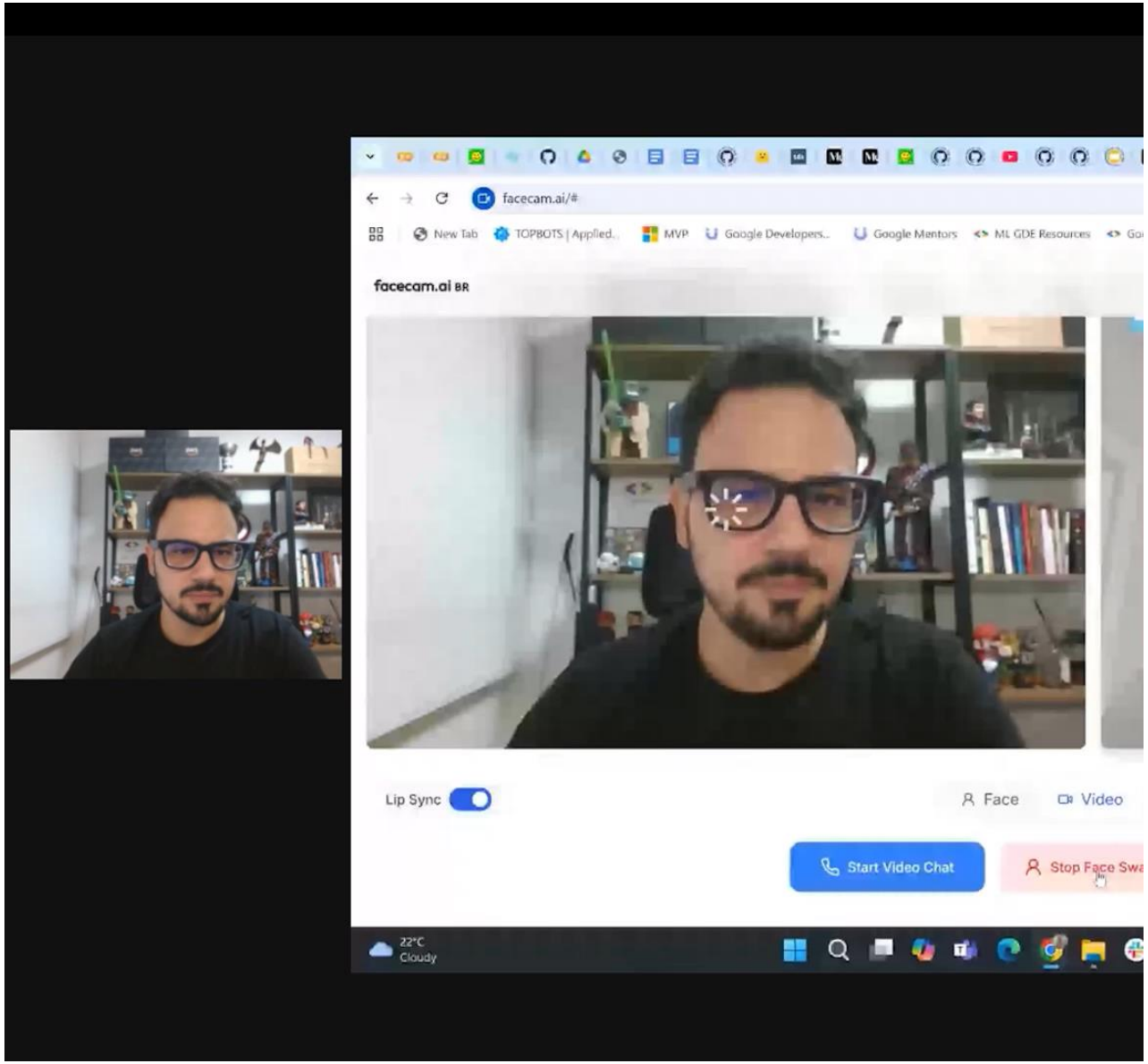

runway


MASTER YARDA
YODDA
VIASTER YODDA
MASTER YODDA
runway


runway

DevFest 2024 Google Developer Groups Cerrado

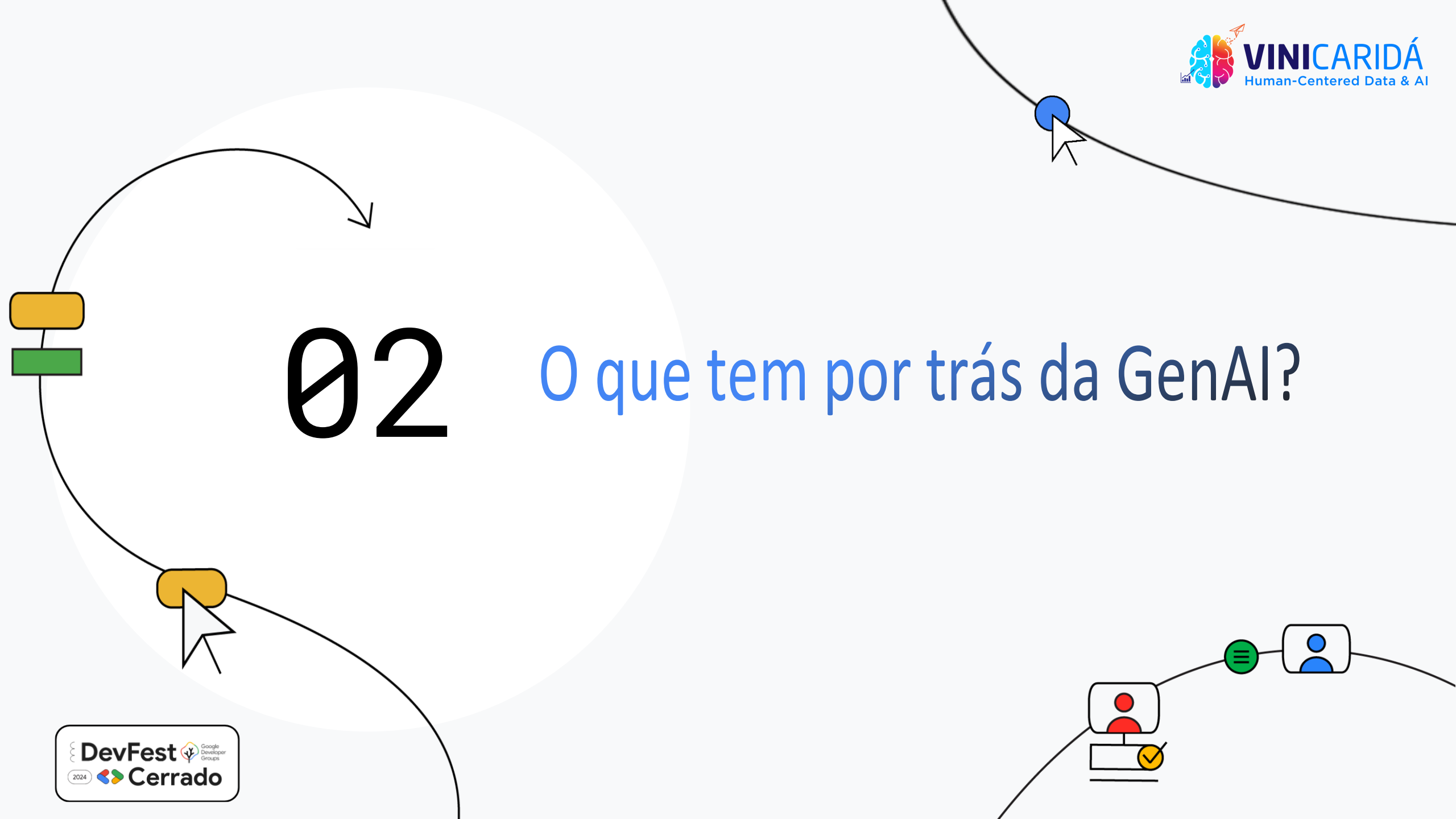
VINICARIDÁ
Human-Centered Data & AI



Experts
Google Developers

VINICARIDÁ
Human-Centered Data & AI

02 O que tem por trás da GenAI?



DevFest 2024 Google Developer Groups Cerrado

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Você sabe qual o significado da palavra **tezgüino**?

exemplo apresentado em: https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf - cap 14

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Consegue entender o que é tezgüino?

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Observe a palavra **tezgüino** em diferentes contextos:

- Uma garrafa de **tezgüino** está sobre a mesa.
- Todo mundo gosta de beber **tezgüino**.
- Você pode ficar bêbado com **tezgüino**.
- **Tezgüino** é feito de milho.

Com o contexto, conseguimos identificar do que se refere a palavra **tezgüino**.

Tezgüino:= é uma bebida alcoólica feita a base de milho.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Como o cérebro faz isso?

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Quais outras palavras se “***encaixam***” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.

Como representar contexto/significado das palavras

Quais outras palavras se “***encaixam***” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Quais outras palavras se “***encaixam***” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Quais outras palavras se “***encaixam***” nos slots das perguntas 1 até 4?

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Inserindo contexto de forma manual...

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

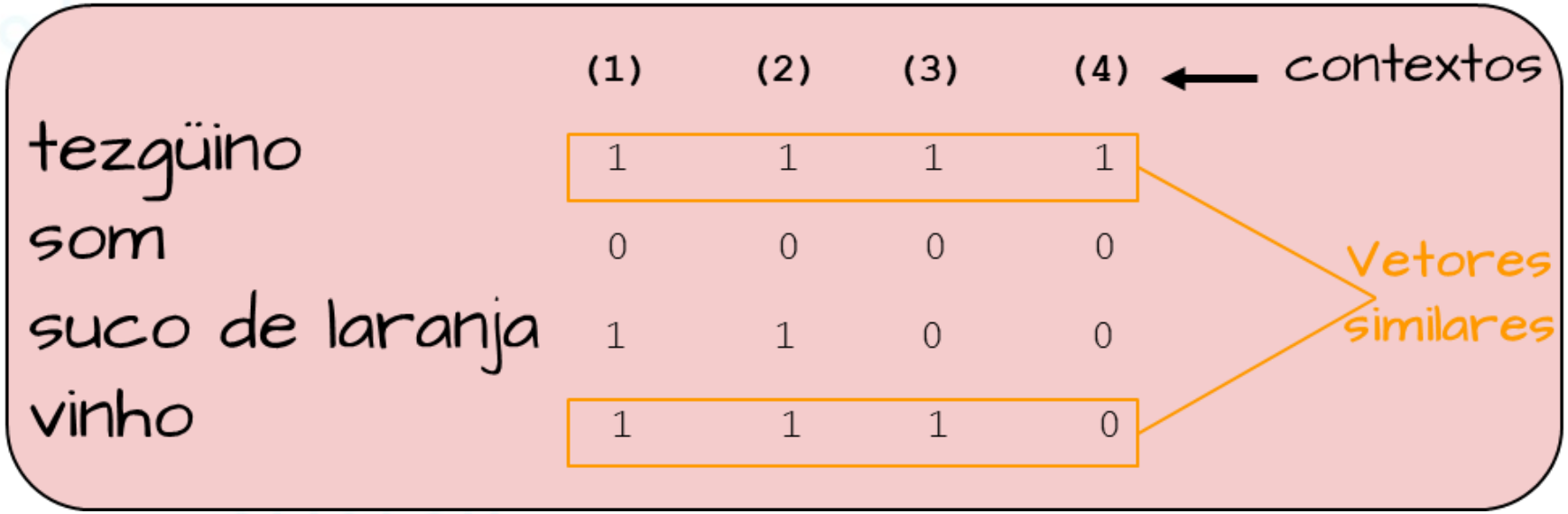
	(1)	(2)	(3)	(4) ← contextos
tezgüino	1	1	1	1
som	0	0	0	0
suco de laranja	1	1	0	0
vinho	1	1	1	0

Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Inserindo contexto de forma manual...

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.



Como representar contexto/significado das palavras

VINICARIDÁ
Human-Centered Data & AI

Inserindo contexto de forma manual...

1. Uma garrafa de _____ está sobre a mesa.
2. Todo mundo gosta de beber _____.
3. Você pode ficar bêbado com _____.
4. _____ é feito de milho.

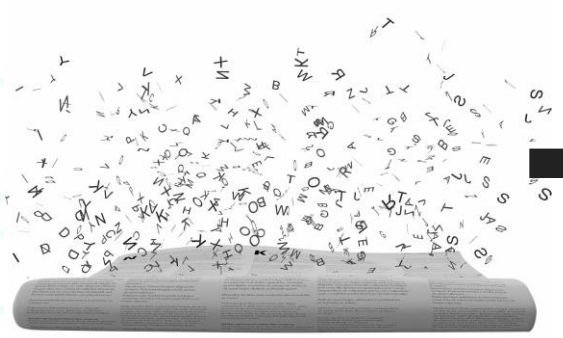
	(1)	(2)	(3)	(4)
tezgüino	1	1	1	1
som	0	0	0	0
suco de laranja	1	1	0	0
vinho	1	1	1	0

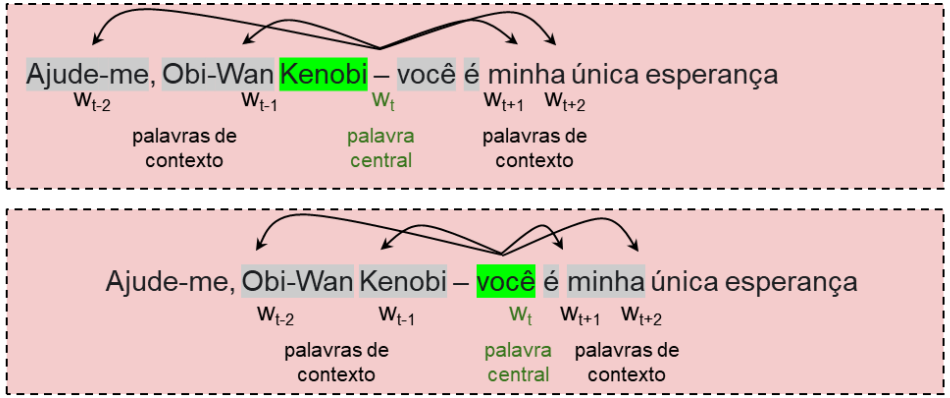
contextos

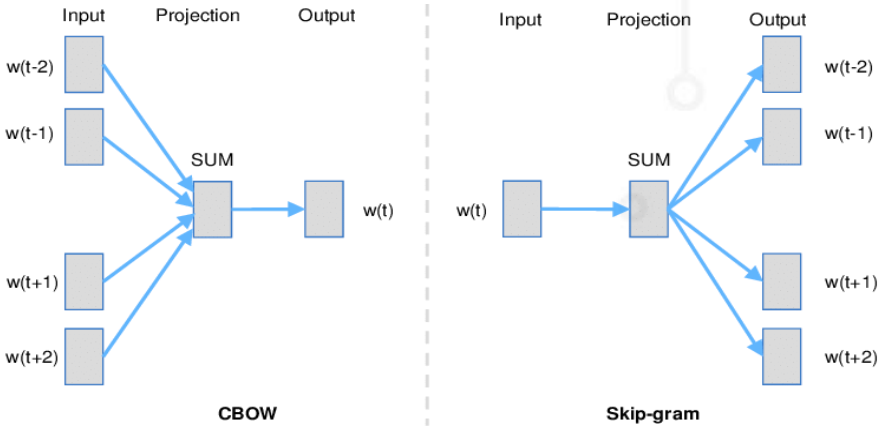
Vetores similares → contexto similar

Processamento de Linguagem Natural

Natural Language Processing (NLP)

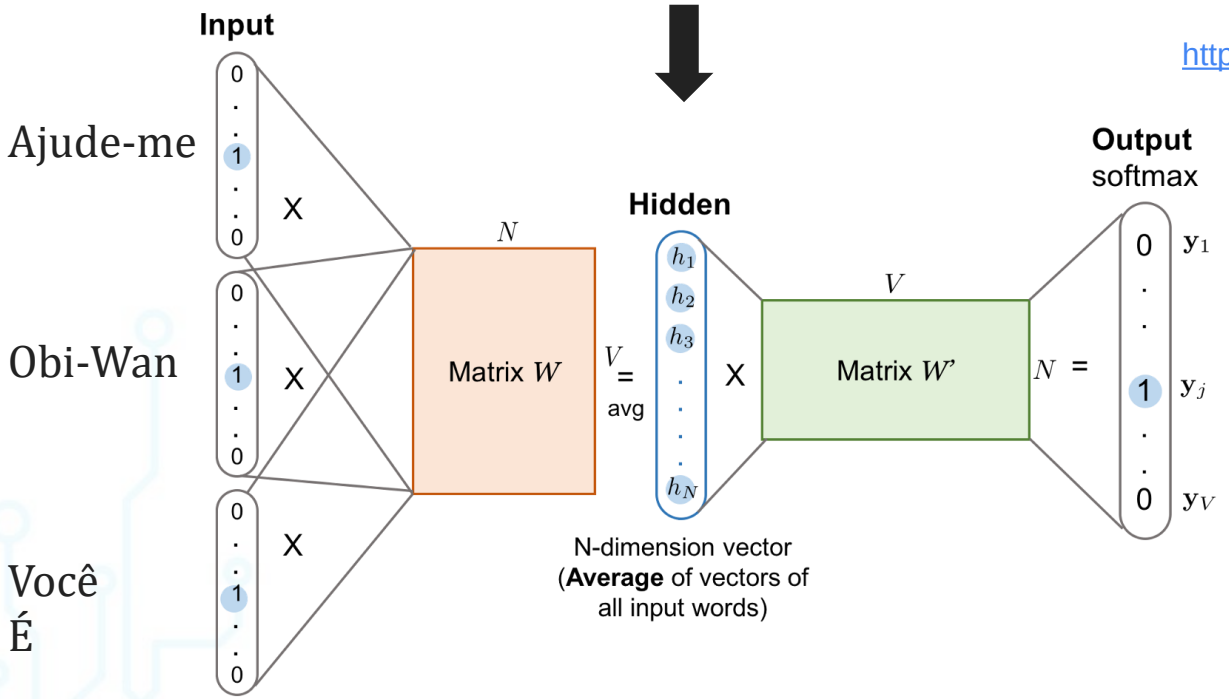






https://arxiv.org/abs/1301.3781

https://arxiv.org/pdf/1310.4546.pdf



MultiLayer Perceptron MLP

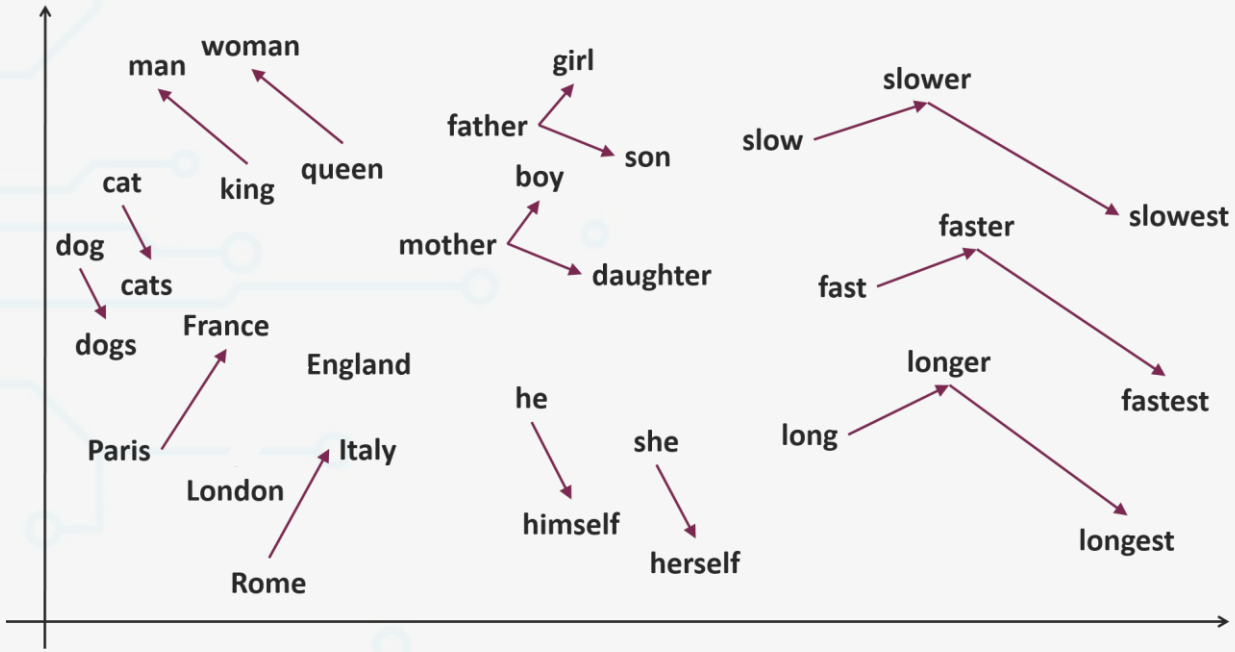
https://dl.acm.org/doi/10.5555/1639537.1639542

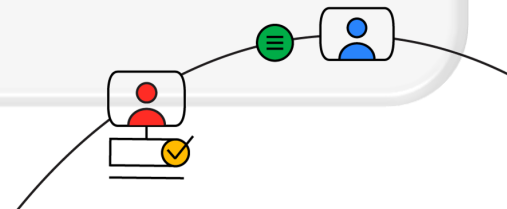
Self Supervised Learning (SSL)

https://arxiv.org/abs/2110.09327

Processamento de Linguagem Natural

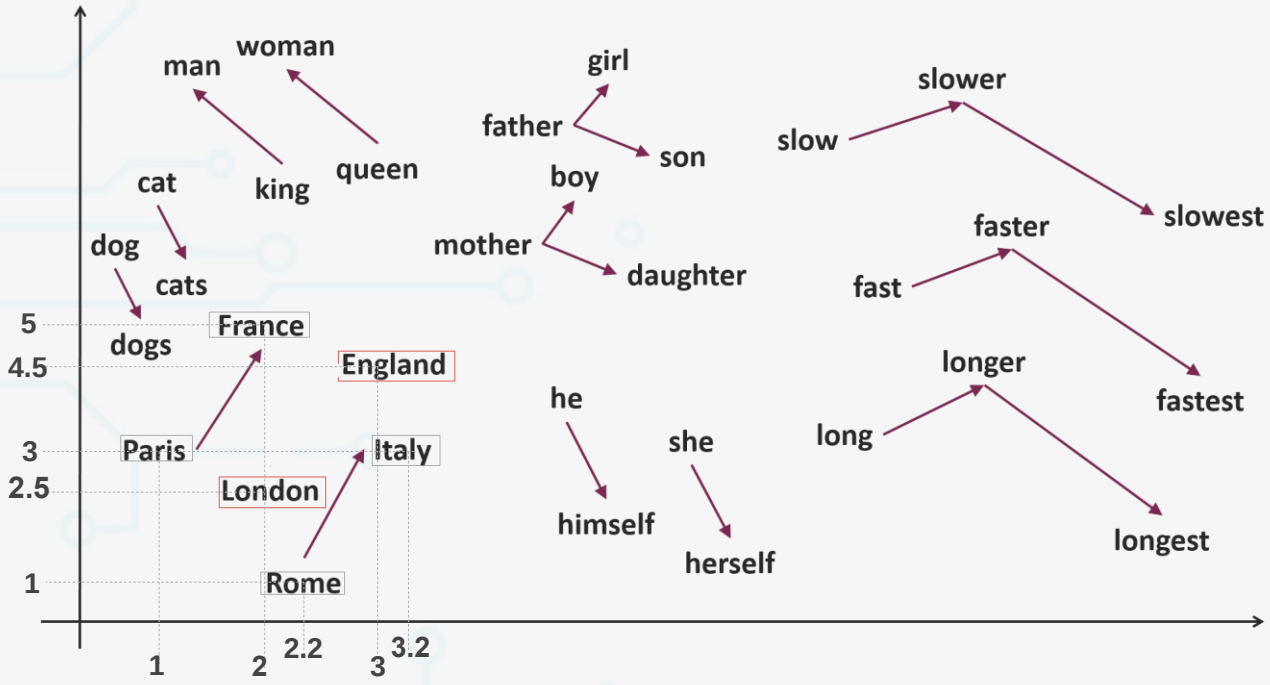
Natural Language Processing (NLP)





Processamento de Linguagem Natural

Natural Language Processing (NLP)



Paris [1, 3]

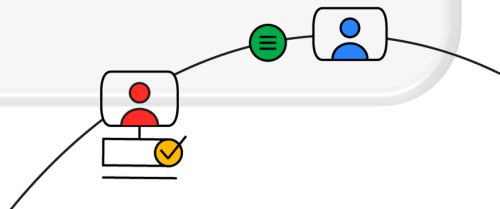
France [2, 5]

London [2, 2.5]

England [3, 4.5]

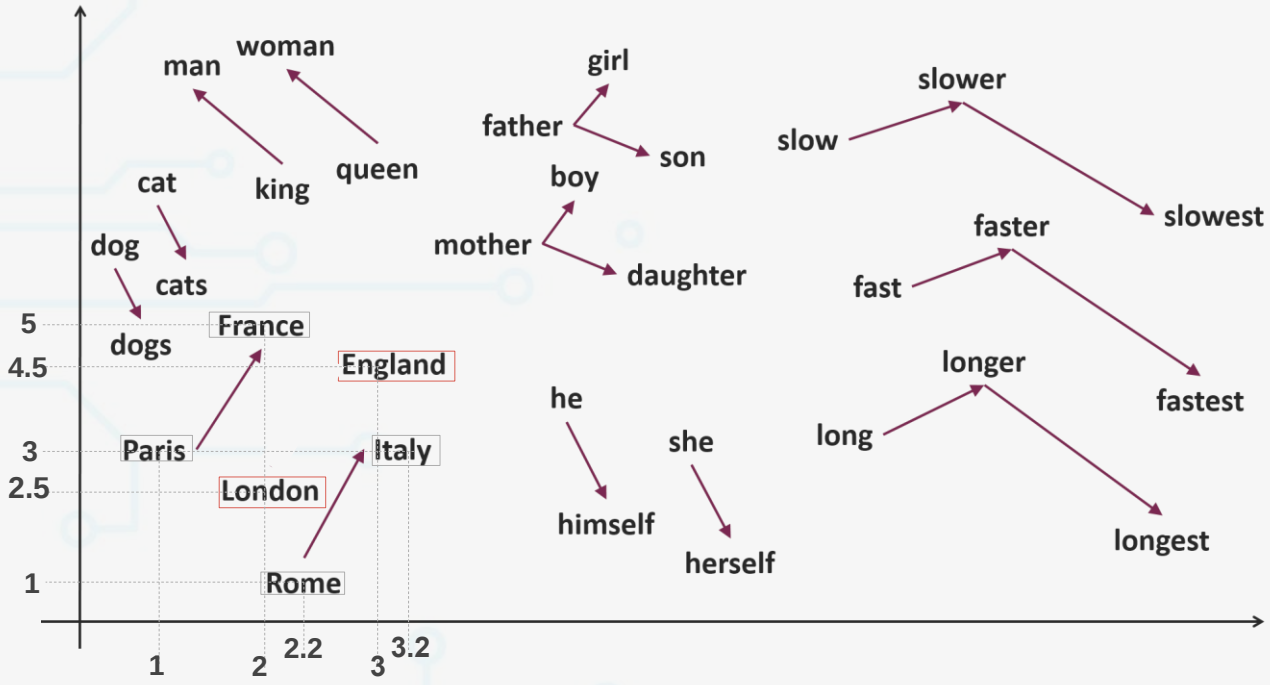
Rome [2.2, 1]

Italy [3.2, 3]

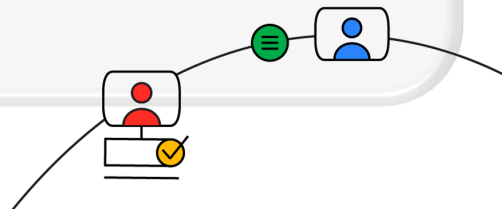


Processamento de Linguagem Natural

Natural Language Processing (NLP)

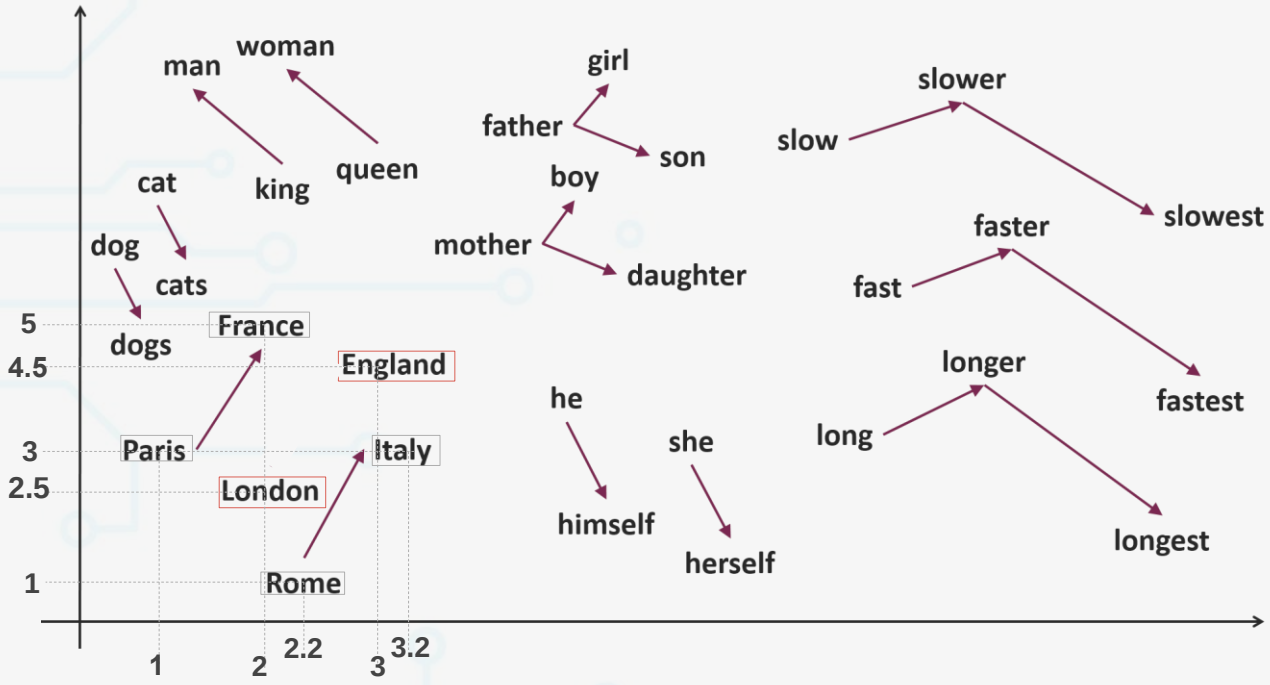


Qual a capital da Inglaterra?



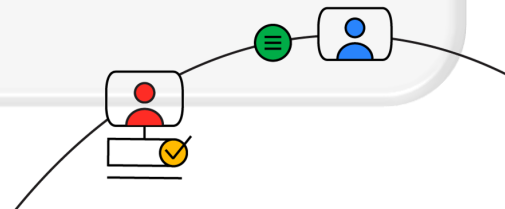
Processamento de Linguagem Natural

Natural Language Processing (NLP)



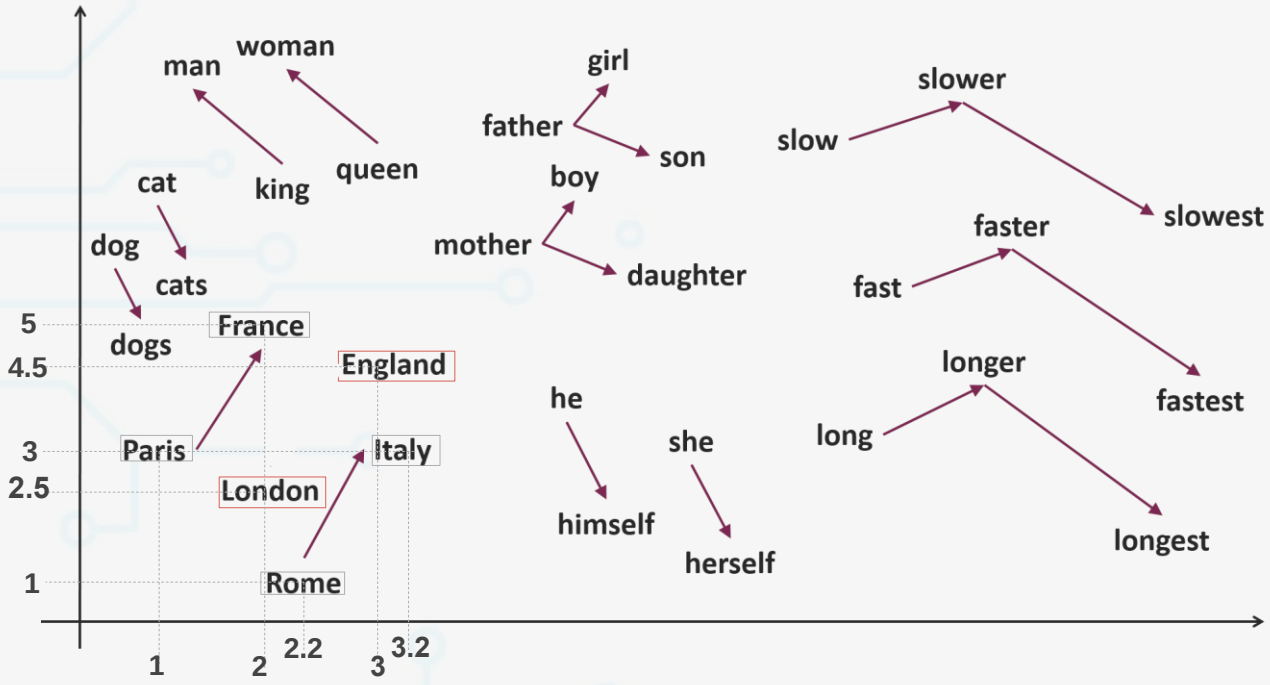
Qual a capital da Inglaterra?

Paris – France + England = ?



Processamento de Linguagem Natural

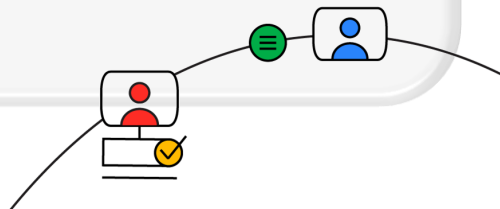
Natural Language Processing (NLP)



Qual a capital da Inglaterra?

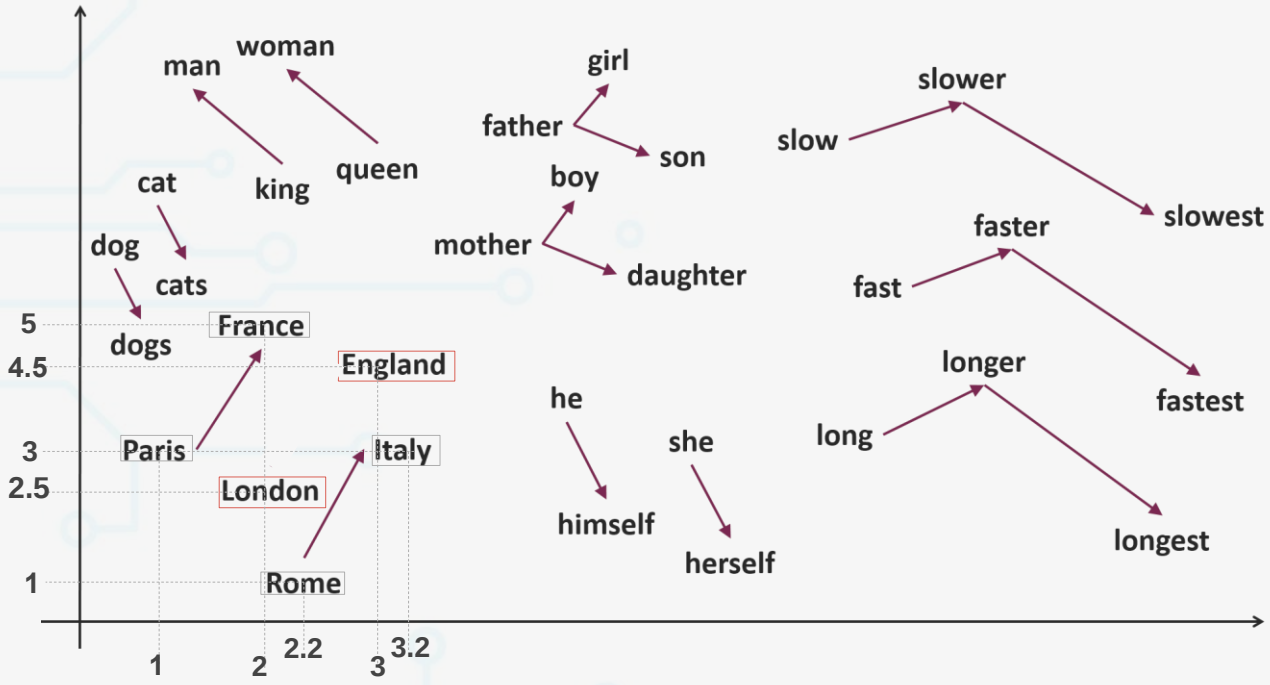
Paris – France + England = ?

Paris [1, 3]
France [2, 5]
=
Result. [-1, -2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]



Processamento de Linguagem Natural

Natural Language Processing (NLP)

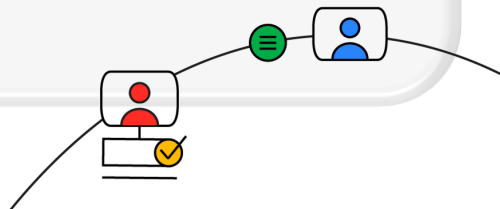


Qual a capital da Inglaterra?

Paris – France + England = ?

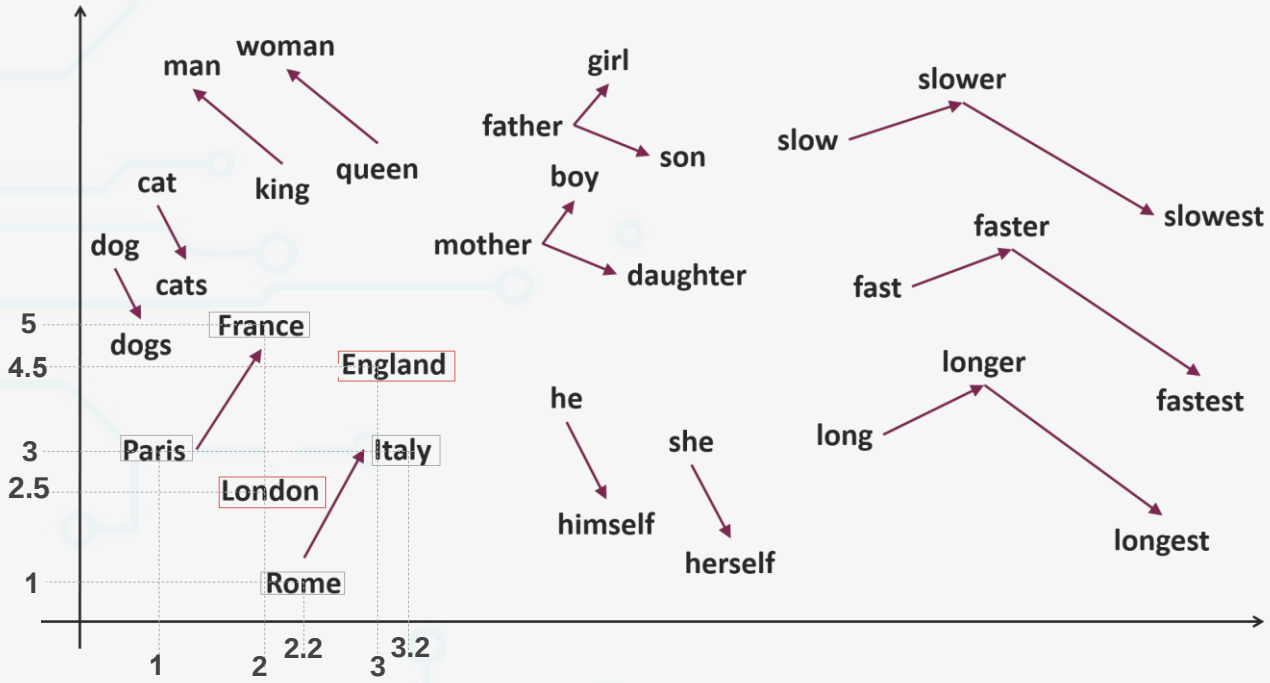
Paris [1, 3]
France [2, 5]
=
Result. [-1, -2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Paris – France + England = London



Processamento de Linguagem Natural

Natural Language Processing (NLP)



Qual a capital da Inglaterra?

Paris – France + England = ?

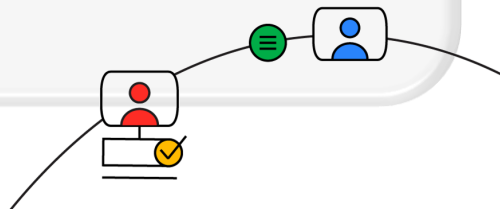
Paris [1, 3]
France [2, 5]
=
Result. [-1, -2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Rome – Italy + England = ?

Rome [2.2, 1]
Italy [3.2, 3]
=
Result. [-1, -2]
+
England [3, 4.5]
=
Result. [2, 2.5]
==
London [2, 2.5]

Paris – France + England = London

Rome – Italy + England = London



Processamento de Linguagem Natural

Natural Language Processing (NLP)

Bag of Words

TF-IDF
Akiko Aizaka, 2003

Word2Vec
Mikolov et al, 2013
Google

Glove
Pennington et al., 2014

Seq2Seq
Sutskever et al, 2014
Google

CoVe
McCann et al., 2017

FastText
Bojanowski et al, 2017
Meta

ULMFiT
Howard et al.. 2018

ELMo
Peters et al, 2018
AI2

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com
Google

https://arxiv.org/abs/1706.03762

DevFest 2024 Google Developer Groups Cerrado

VINICARIDÁ Human-Centered Data & AI

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

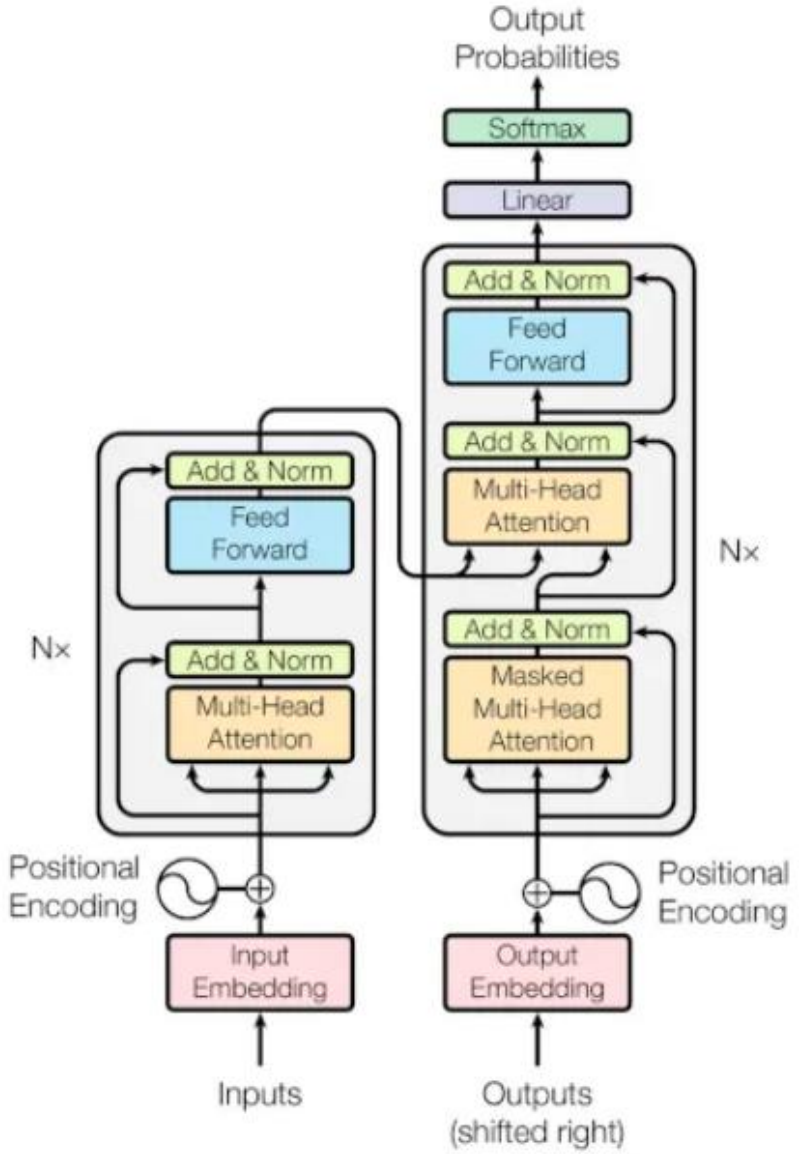
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

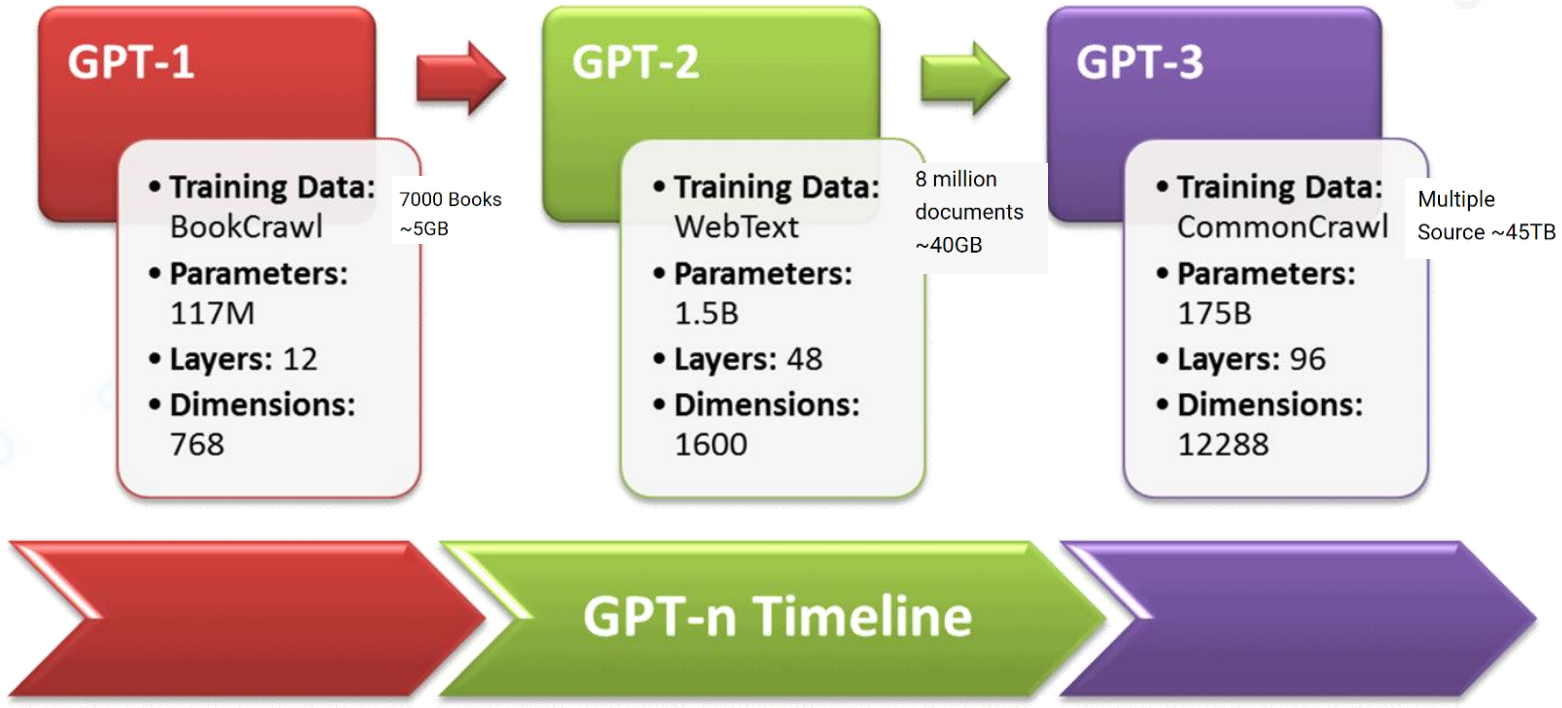
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task,

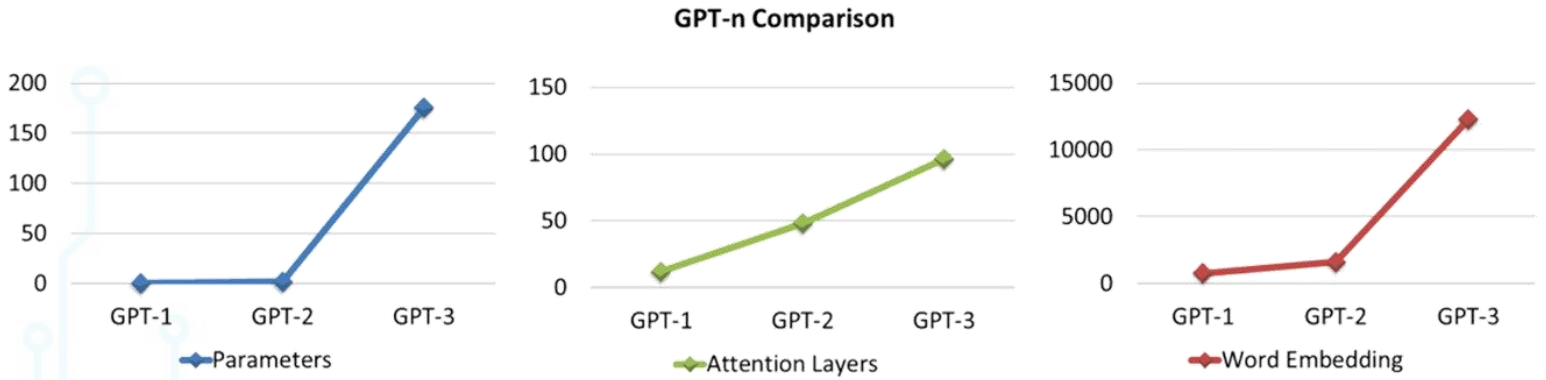
https://arxiv.org/abs/1706.03762



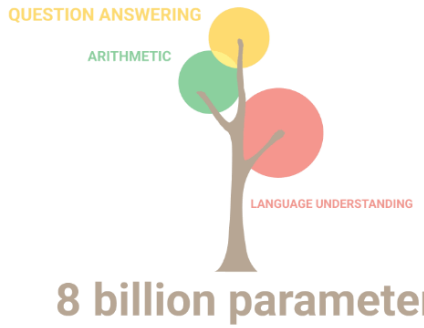
Experts Google Developers

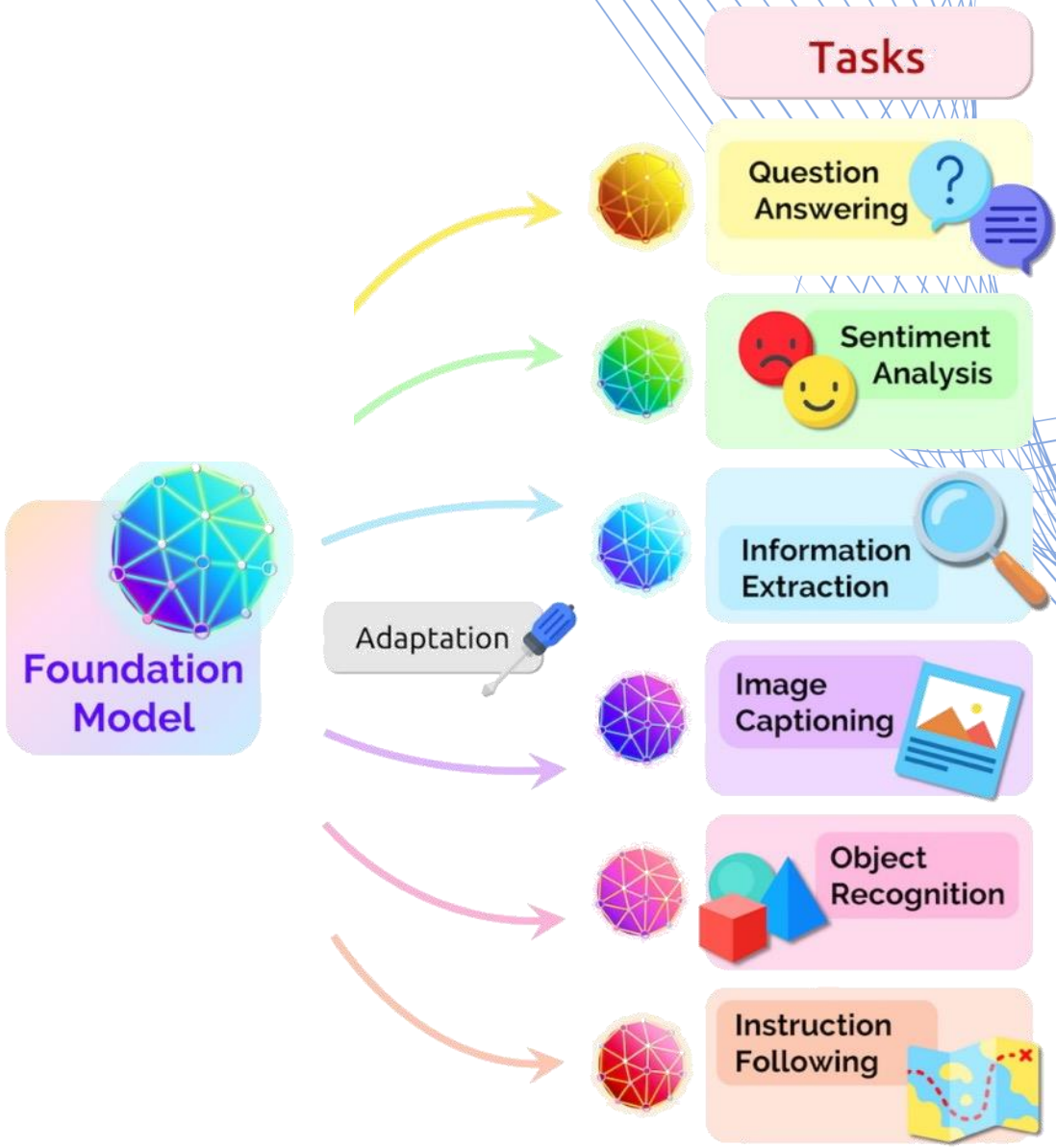
Família GPT





https://businessolution.org/gpt-3-statistics/





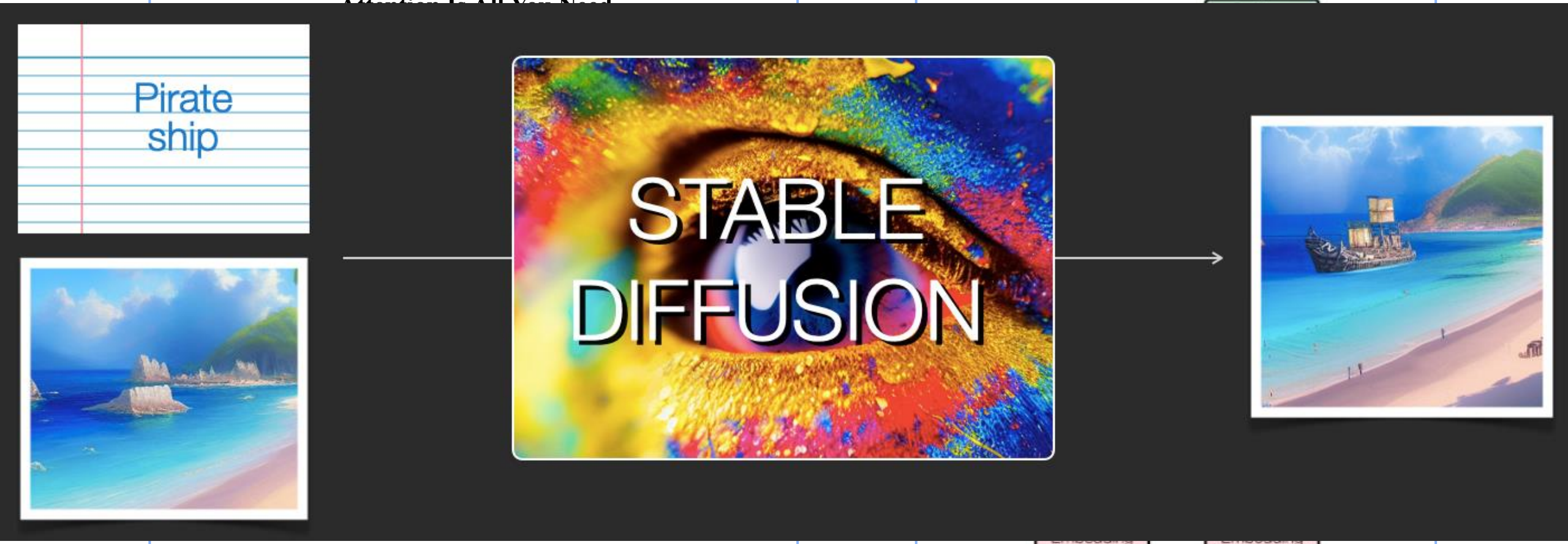
Zero-Shot Learning in Modern NLP
https://joeddav.github.io/blog/2020/05/29/ZSL.html

Few-Shot Learning
Não Corporativo https://arxiv.org/pdf/1904.05046.pdf

DevFest 2024 Cerrado

VINICARIDÁ
Human-Centered Data & AI

Output Probabilities



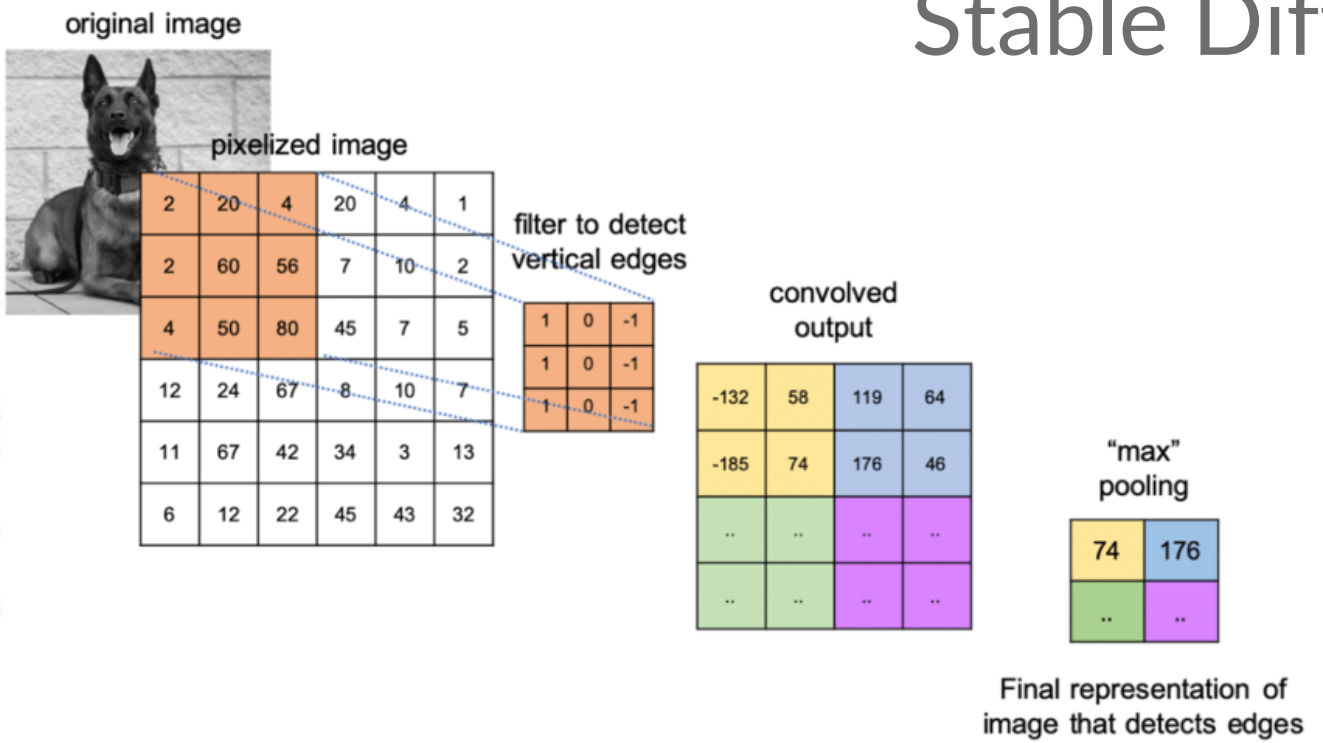
Inputs

Outputs (shifted right)

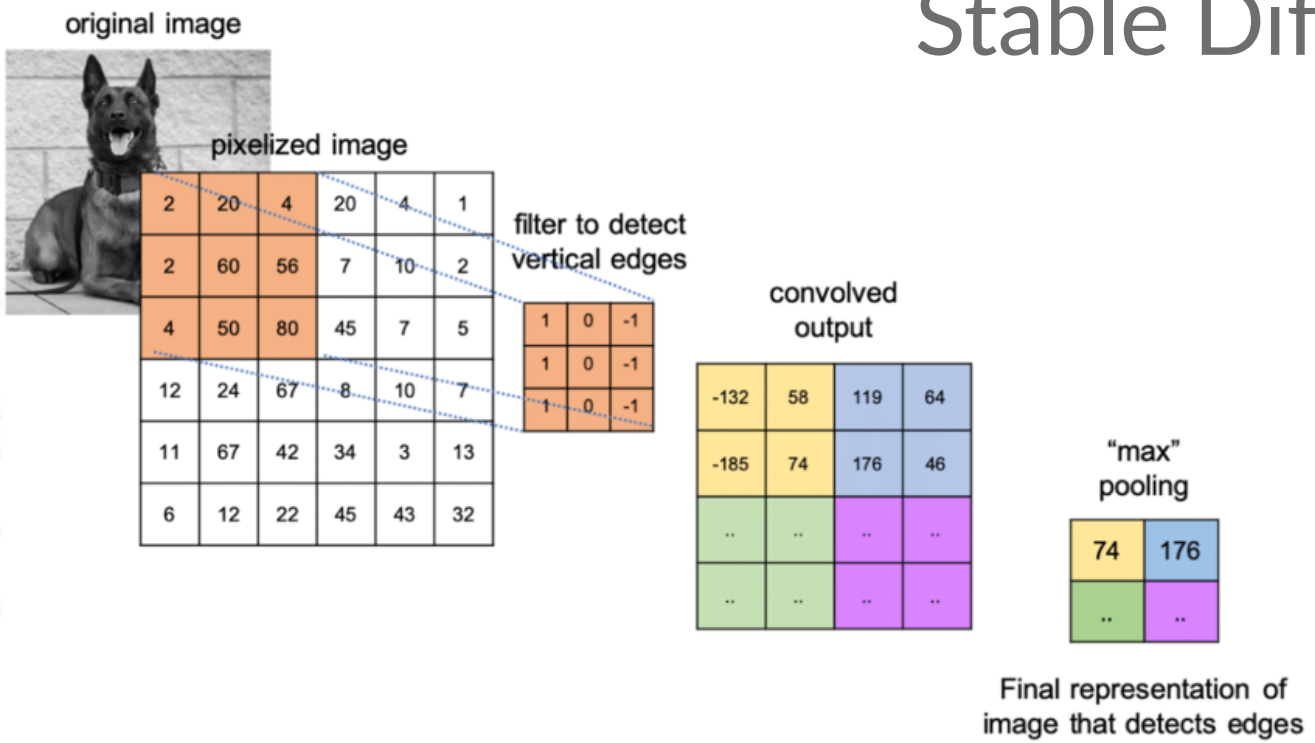
https://arxiv.org/abs/1706.03762

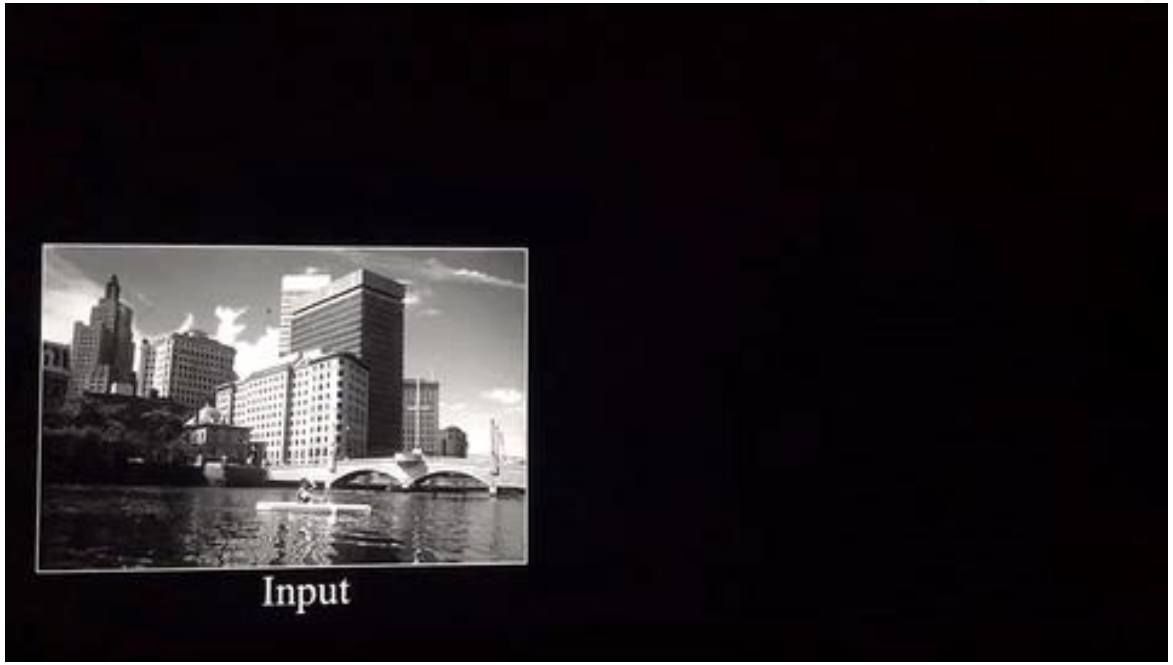
Experts
Google Developers

Stable Diffusion

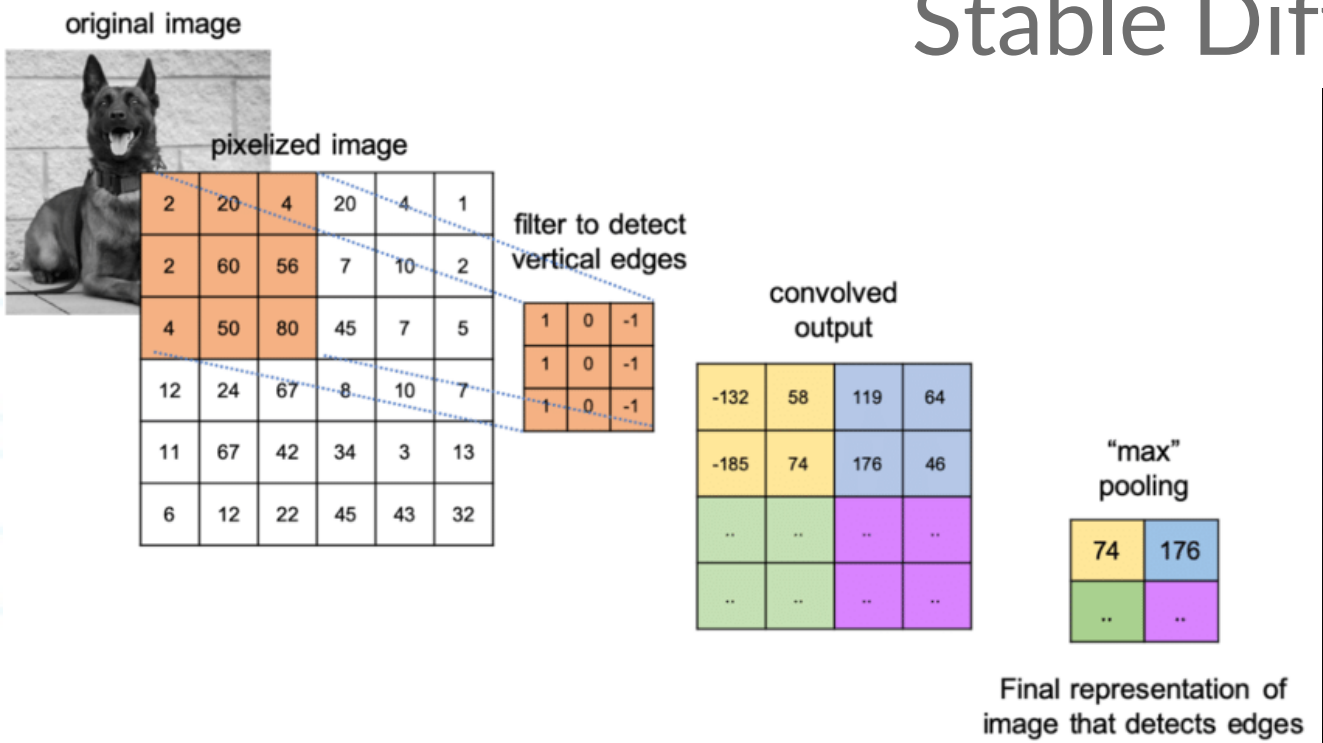


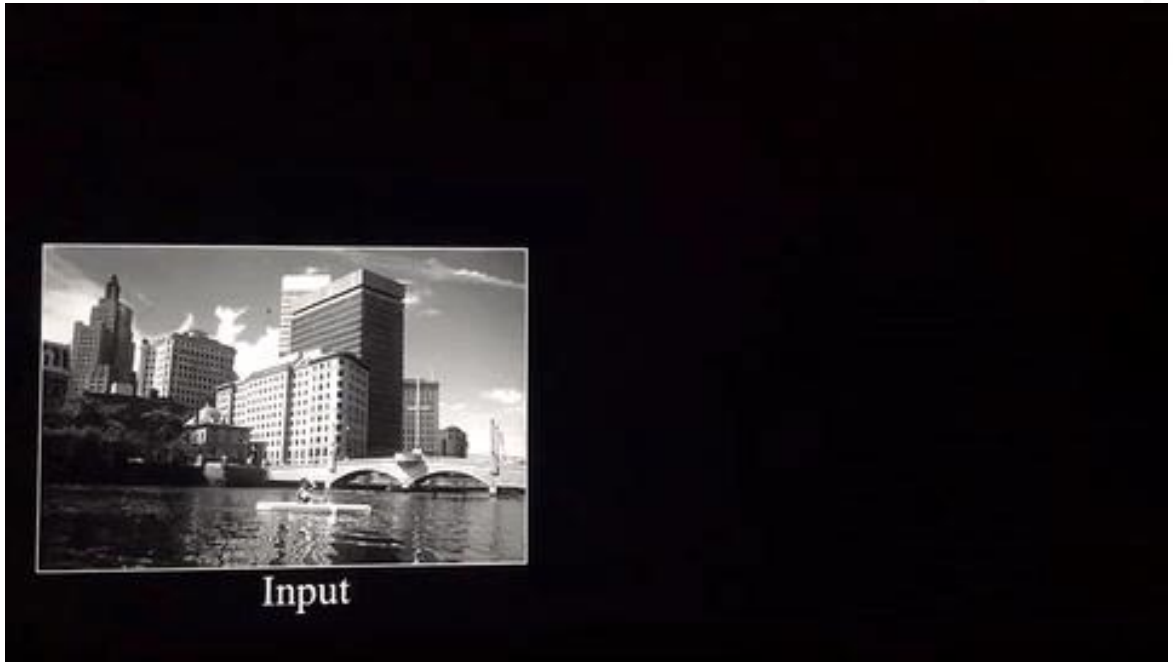
Stable Diffusion





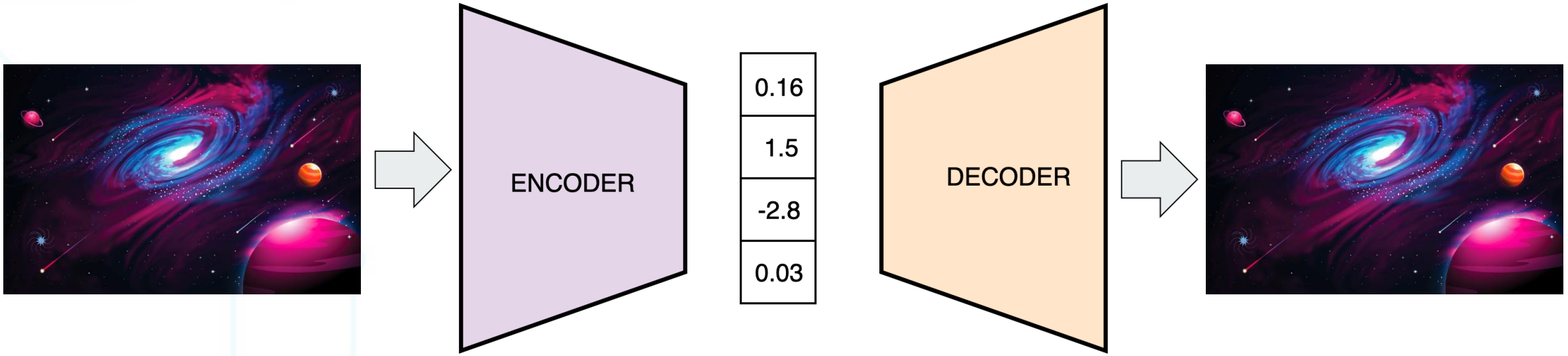
Stable Diffusion





Stable Diffusion

Espaço latente



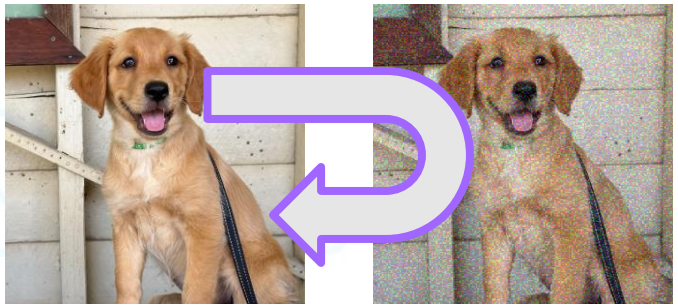
Modelos de difusão

Diffusion model

Adds noise and learns how to work backwards to the original image.

Modelos de difusão

Diffusion model



Adds noise and learns how to work backwards to the original image.

Modelos de difusão

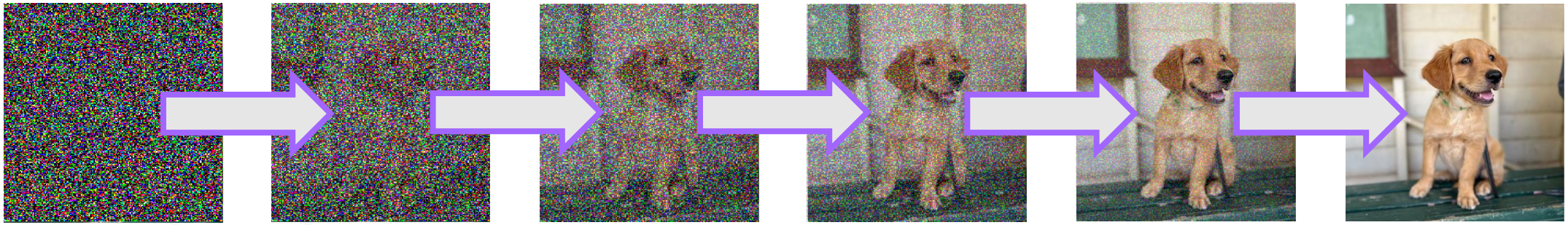
Diffusion model

Adds noise and learns how to work backwards to the original image.

Modelos de difusão

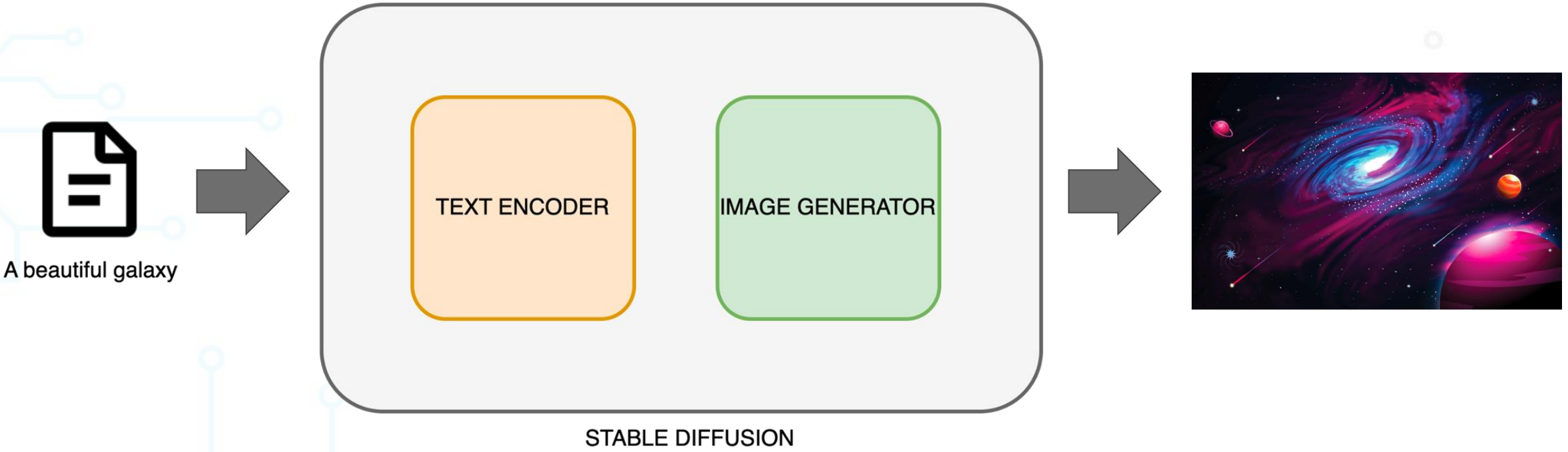
Diffusion model

Adds noise and learns how to work backwards to the original image.



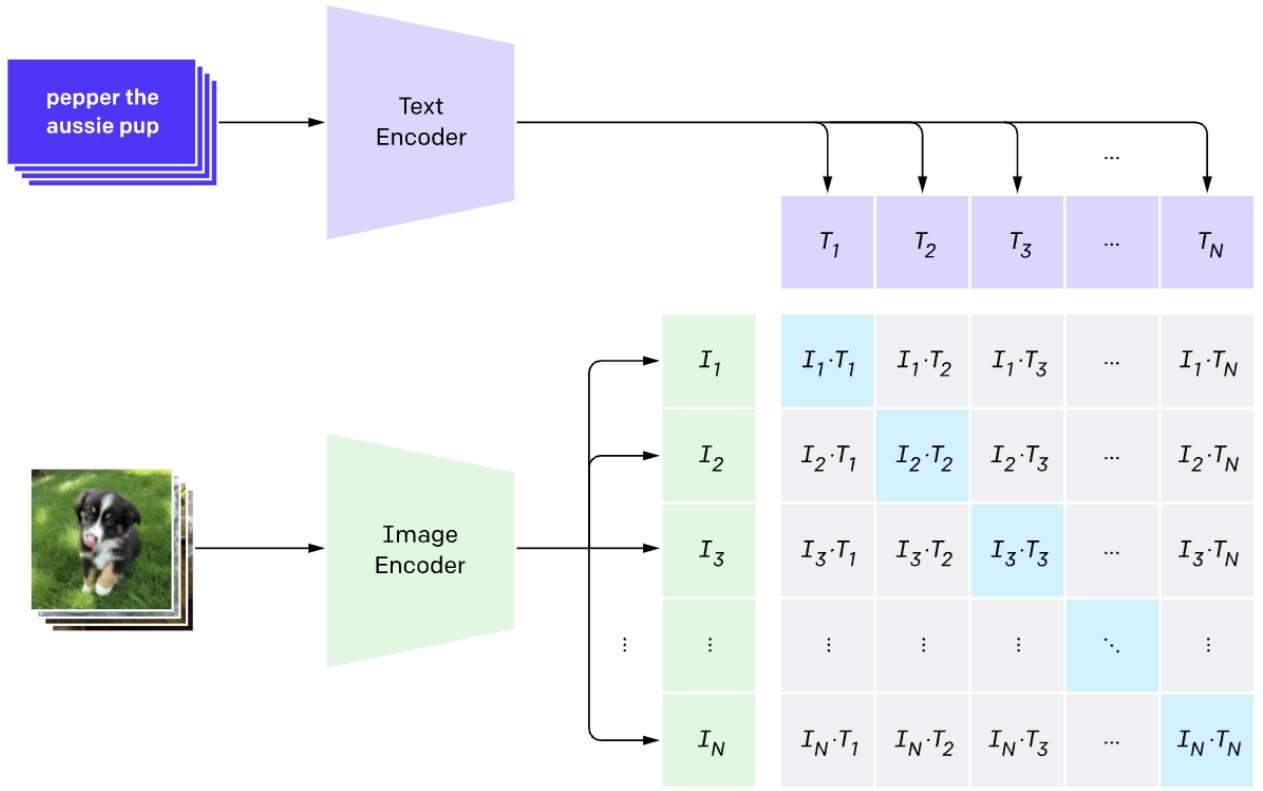
Trained model works from random noise to generated an image.

Stable Diffusion



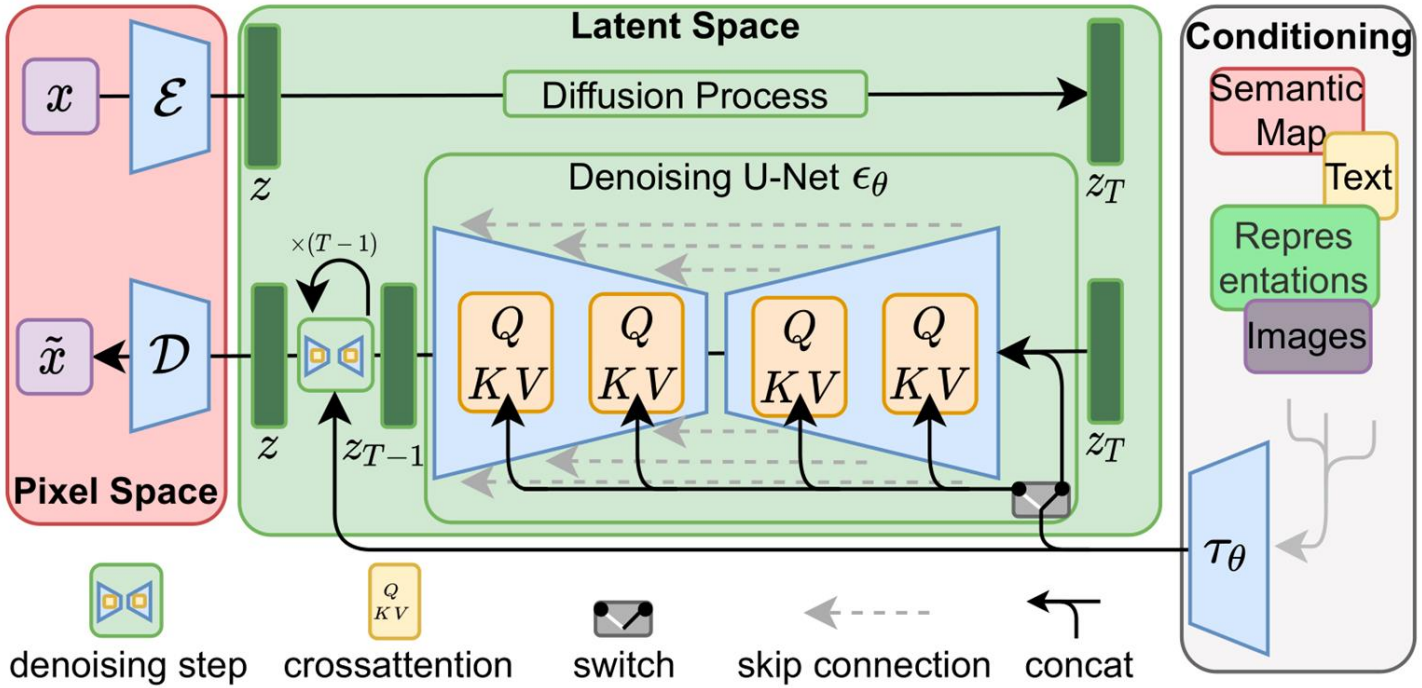
Stable Diffusion

Text Encoder - CLIP



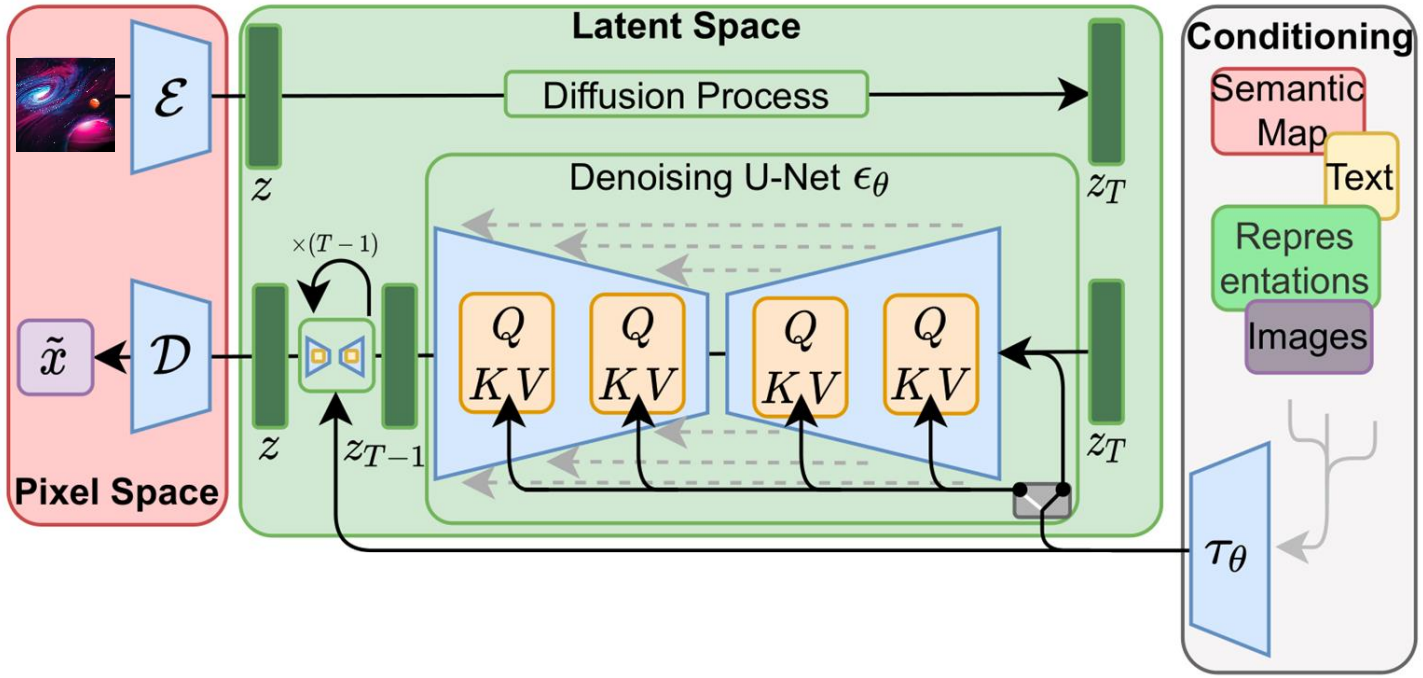
Stable Diffusion

Treinamento



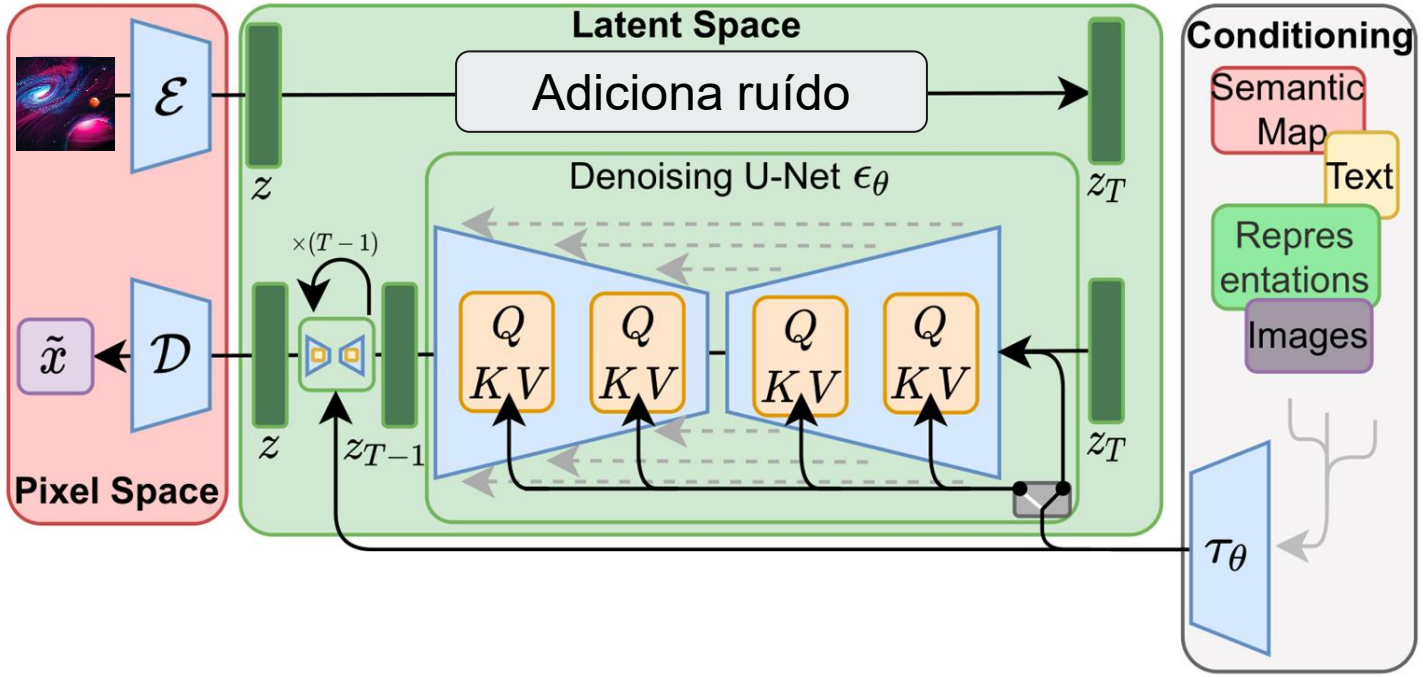
Stable Diffusion

Treinamento



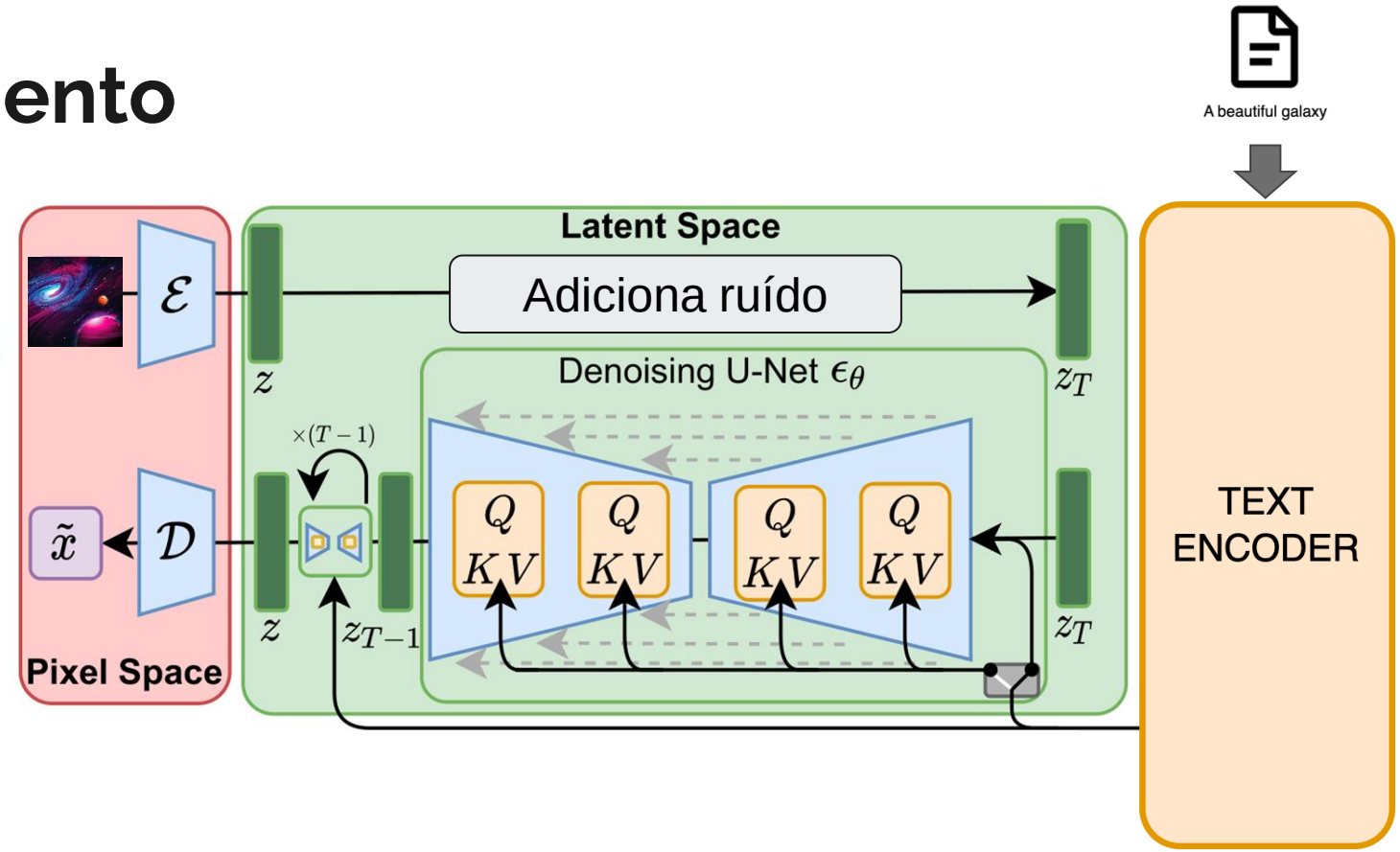
Stable Diffusion

Treinamento



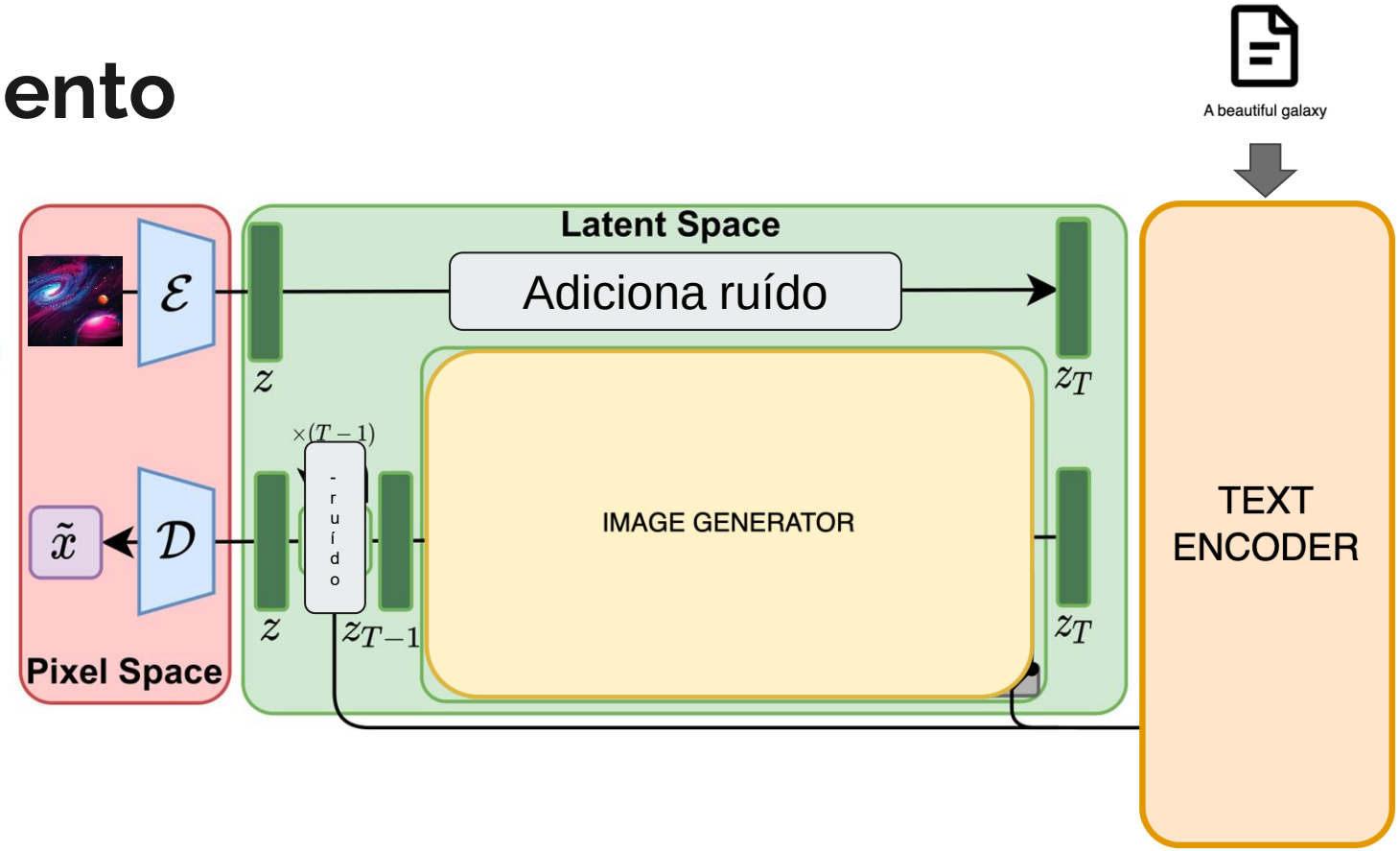
Stable Diffusion

Treinamento



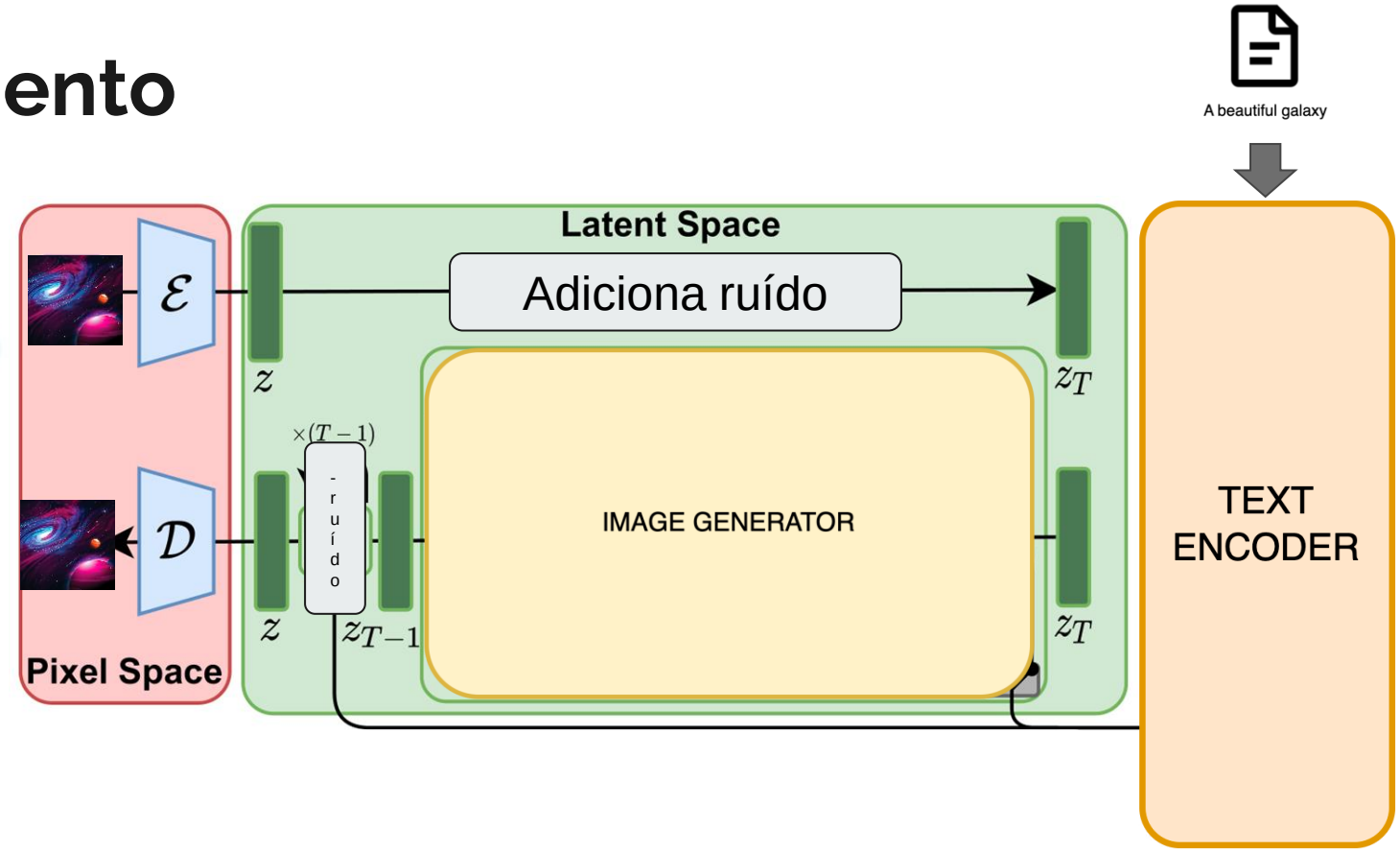
Stable Diffusion

Treinamento

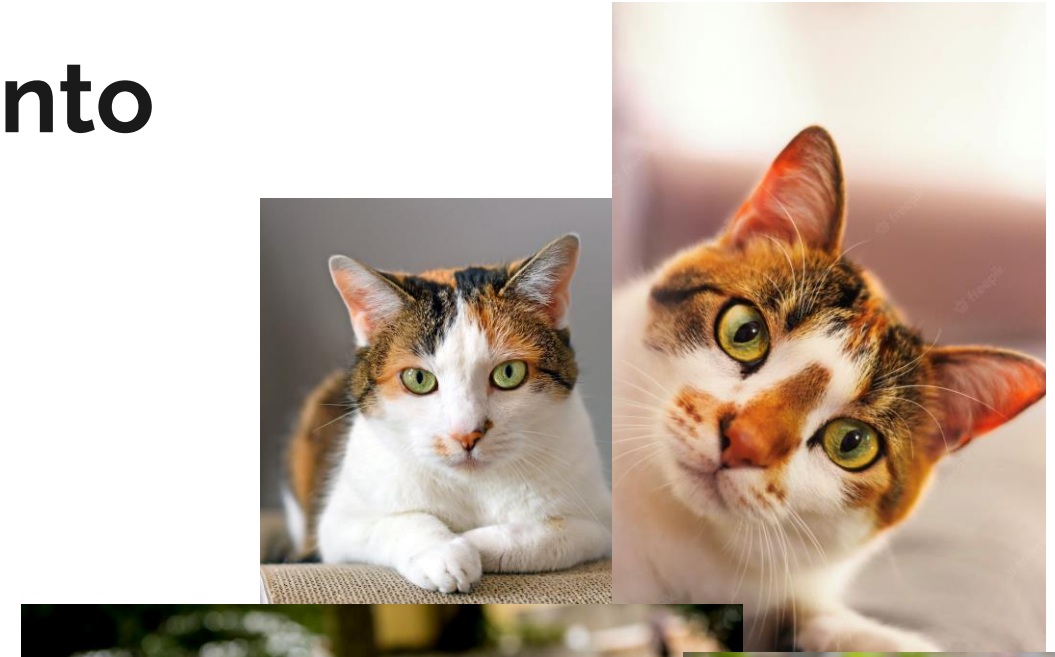


Stable Diffusion

Treinamento



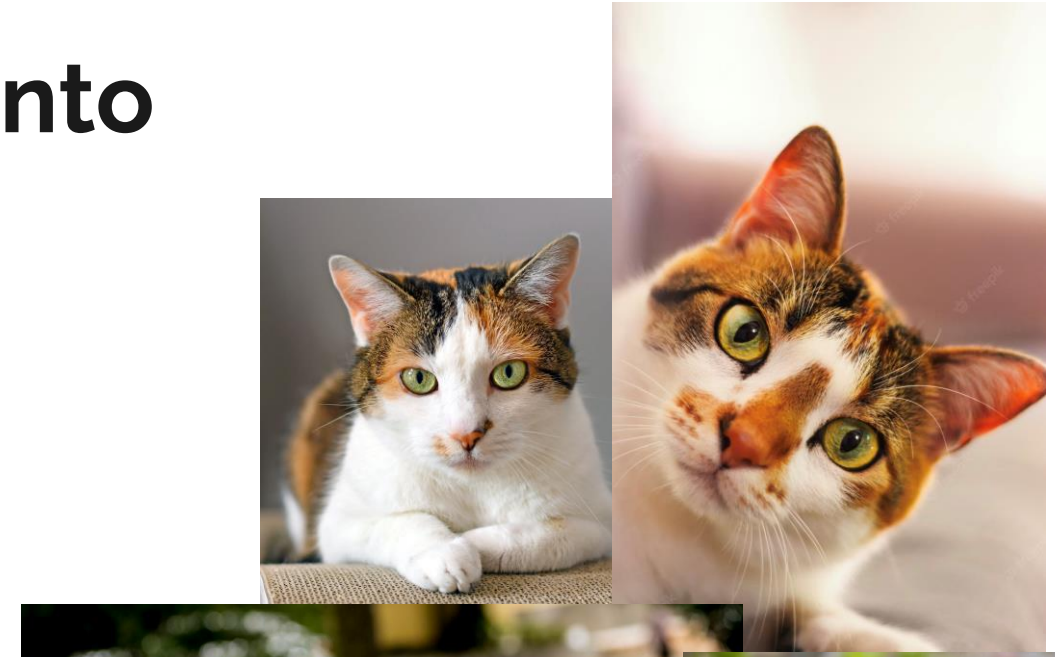
Dados fora do treinamento



Textual Inversion

Dados fora do treinamento

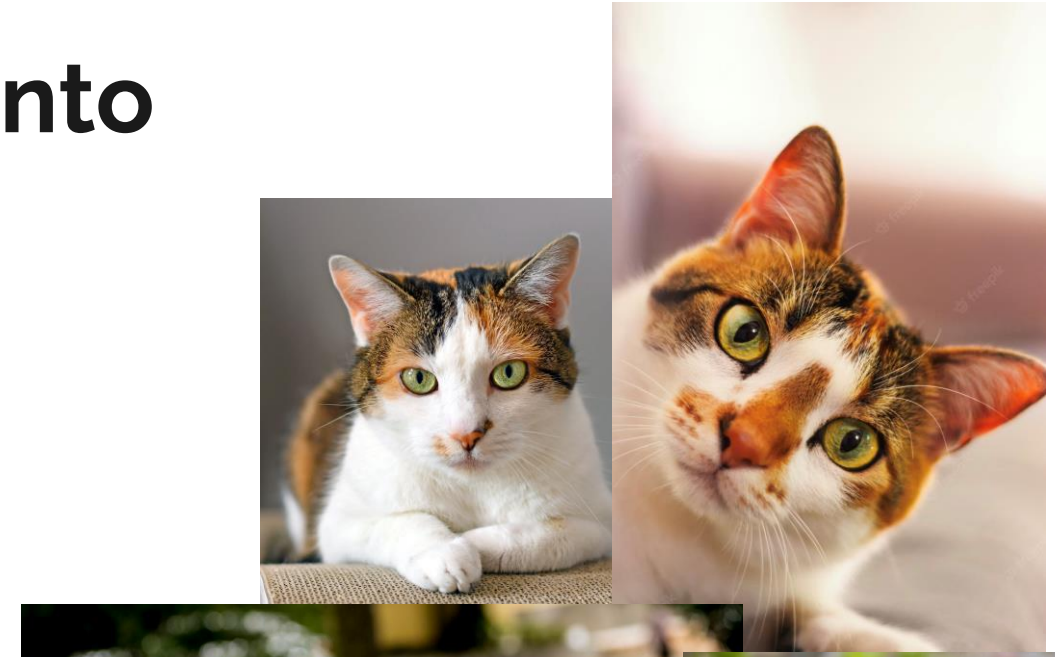
- Um gato em uma praça ✅



Textual Inversion

Dados fora do treinamento

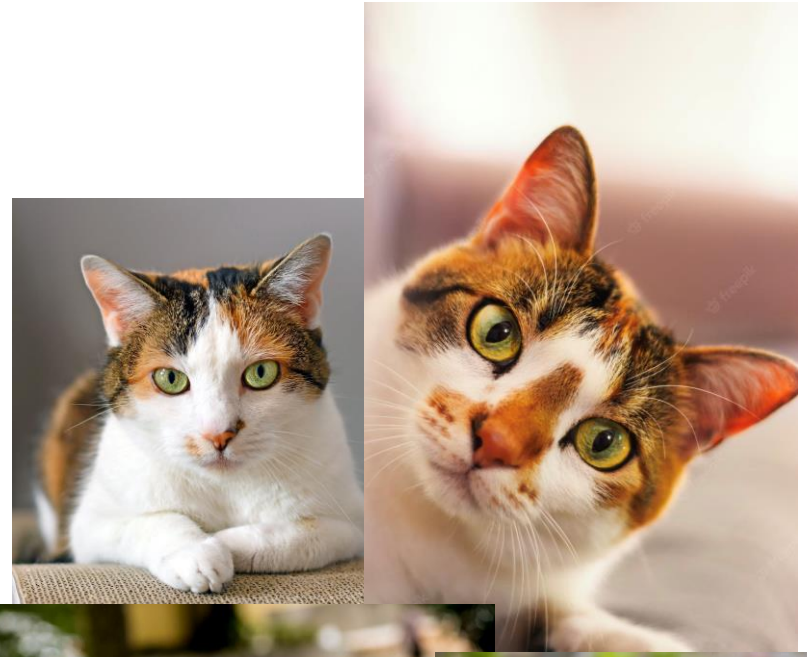
- Um gato em uma praça ✅
- Um cachorro na grama ✅



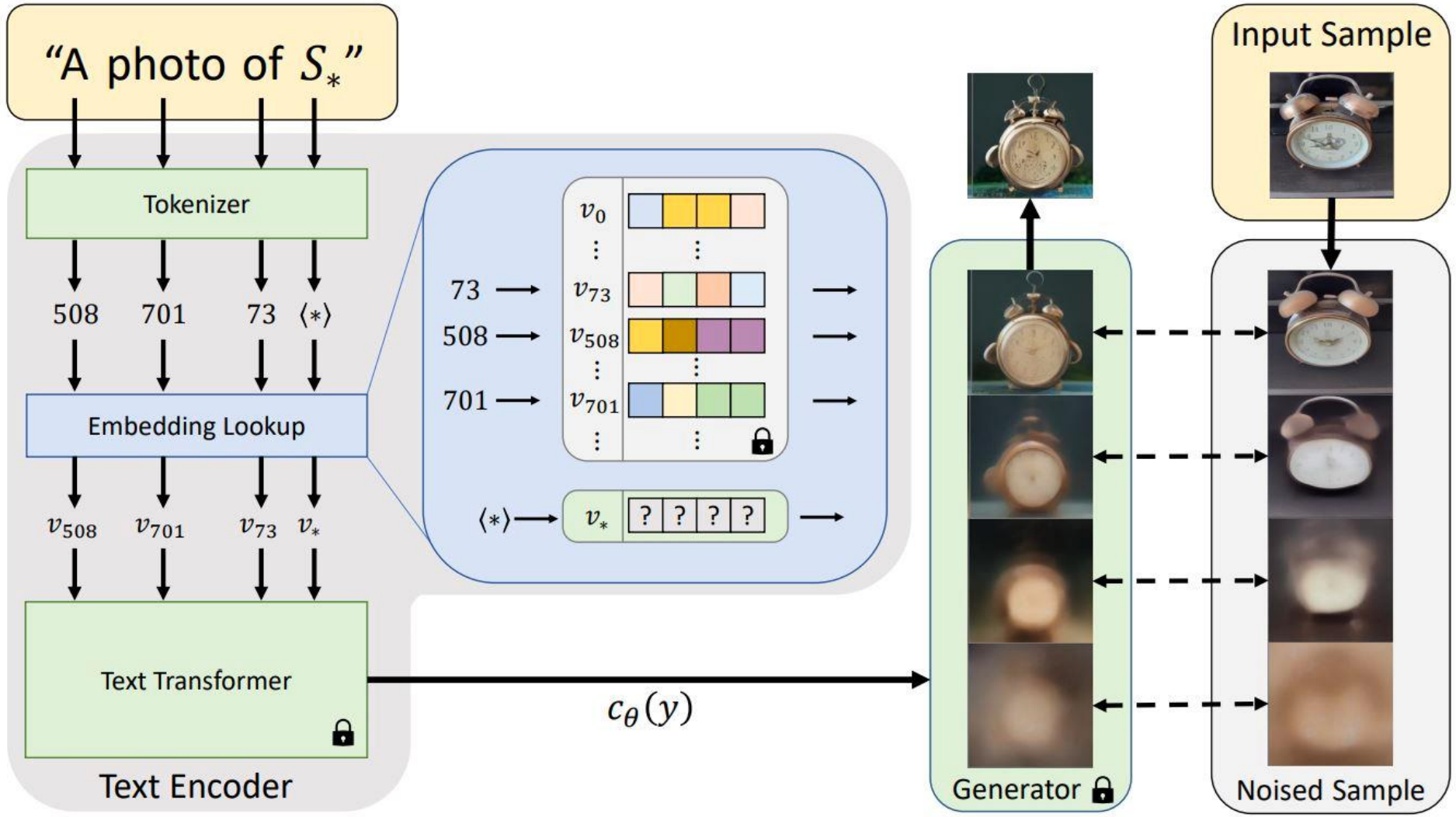
Textual Inversion

Dados fora do treinamento

- Um gato em uma praça ✅
- Um cachorro na grama ✅
- Um coelho no mato ✖



Textual Inversion



VINICARIDÁ
Human-Centered Data & AI

03 Onde o Google entra nisso?

DevFest 2024 Google Developer Groups Cerrado

Google has pioneered AI advances for a decade

VINICARIDÁ
Human-Centered Data & AI

2015

Google DeepMind AlphaGo defeats Go champion

2016

Google's DeepMind helps detect eye disease

2017

Google invents Transformer kickstarting LLM revolution

2018

Google's groundbreaking large language model, BERT

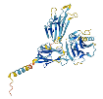
T5

2019

Text-to-Text Transfer Transformer LLM 10B P Model Open Sourced

2020

Google LaMDA Model Trained to converse



2022

AlphaFold predicts structures of all known proteins

2023

A conversational AI Service powered by PaLM2

2024

Family of multimodal LLMs & products

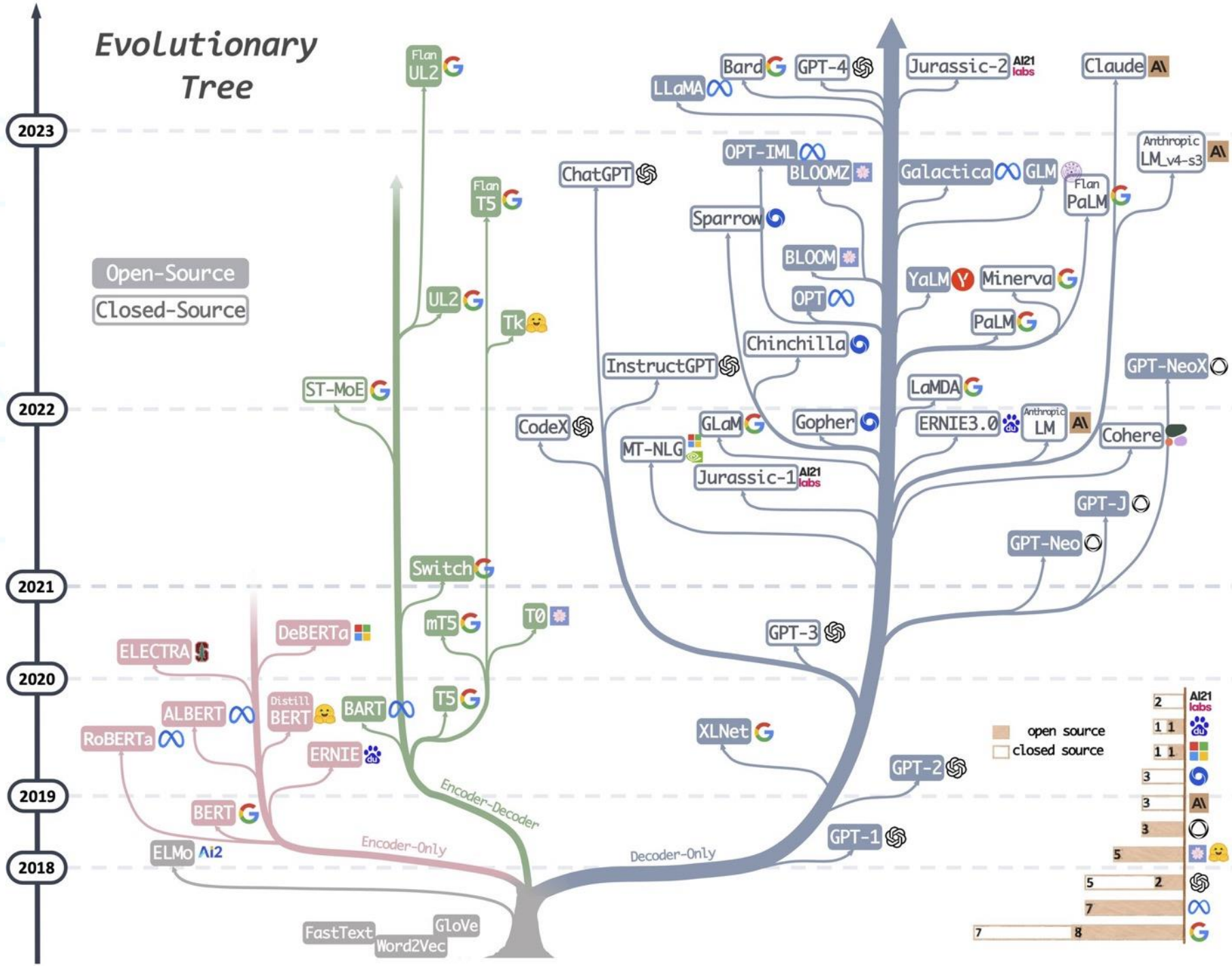
3,000
Researchers

7,000
Publications

Responsible AI

- Built & Tested for Safety
- Upholds high scientific standards
- Socially Beneficial
- Privacy in design
- Accountable to People
- Avoid creating unfair bias

DevFest 2024 Google Developer Groups Cerrado


Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Flan UL2
Flan T5
UL2
Tk
ST-MoE
Switch
mT5
T0
T5
BART
ELECTRA
DeBERTa
ALBERT
Distill BERT
RoBERTa
ERNIE
BERT
ELMo
Ai2
FastText
GloVe
Word2Vec
Encoder-Only
Encoder-Decoder
Decoder-Only
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
ChatGPT
Galactica
GLM
Flan PaLM
Anthropic LM_v4-s3
Sparrow
BLOOM
YaLM
Minerva
OPT
PaLM
Chinchilla
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GPT-Neo
GPT-3
XLNet
GPT-2
GPT-1
open source
closed source
2
1 1
1 1
3
3
3
5
5 2
7
7 8

DevFest 2024 Cerrado

VINICARIDÁ Human-Centered Data & AI

Linha do tempo até o chat Generative Pre-Training chatGPT

OpenAI

2017 — Deep RL from human preferences (link)

2018 — GPT1

2019 — GPT2

2019 — GPT2 treino RLHF (link)

2020 — GPT3

2021 — WebGPT GPT3 fine-tuned

2022 — GPT3.5 GPT3 com RLHF instructGPT

2023 — chatGPT InstructGPT

Deep Reinforcement Learning from Human Preferences

Paul F Christiano
OpenAI
paul@openai.com

Jan Leike
DeepMind
leike@google.com

Tom B Brown
nottombrown@gmail.com

Miljan Martic
DeepMind
miljanm@google.com

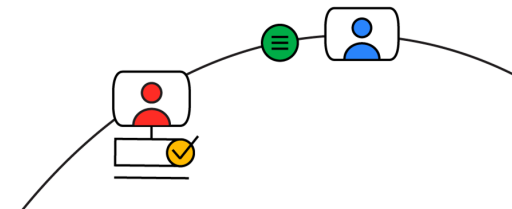
Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com

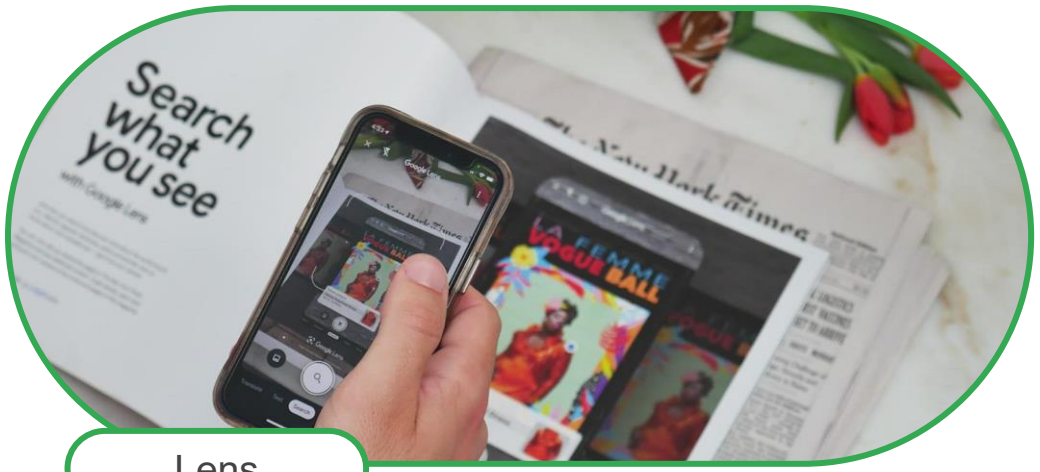
Abstract

For sophisticated reinforcement learning (RL) systems to interact usefully with real-world environments, we need to communicate complex goals to these systems. In this work, we explore goals defined in terms of (non-expert) human preferences between pairs of trajectory segments. We show that this approach can effectively solve complex RL tasks without access to the reward function, including Atari games and simulated robot locomotion, while providing feedback on less than

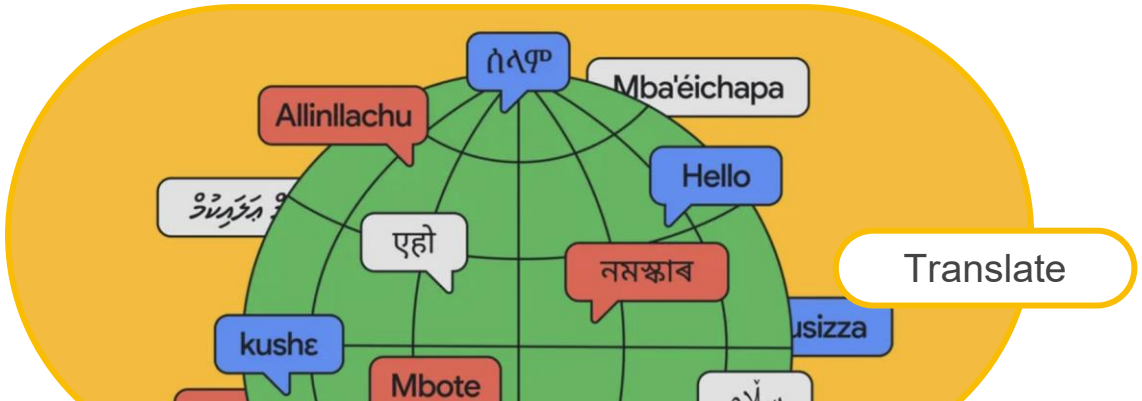
Experts Google Developers



LaMDA

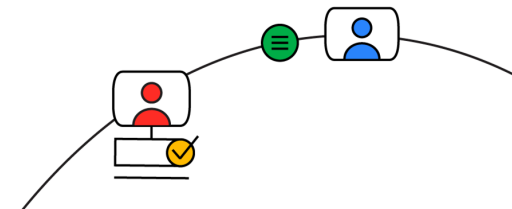


Lens



Translate

Ultimately helping people achieve their potential.



DevFest 2024 Cerrado

VINICARIDÁ Human-Centered Data & AI

Google’s AI Principles

1. Be socially beneficial.

2. Avoid creating or reinforcing unfair bias.

3. Be built and tested for safety.

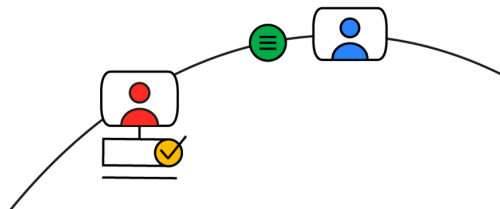
4. Be accountable to people.

5. Incorporate privacy design principles.

6. Uphold high standards of scientific excellence.

7. Be made available for uses that accord with these principles.

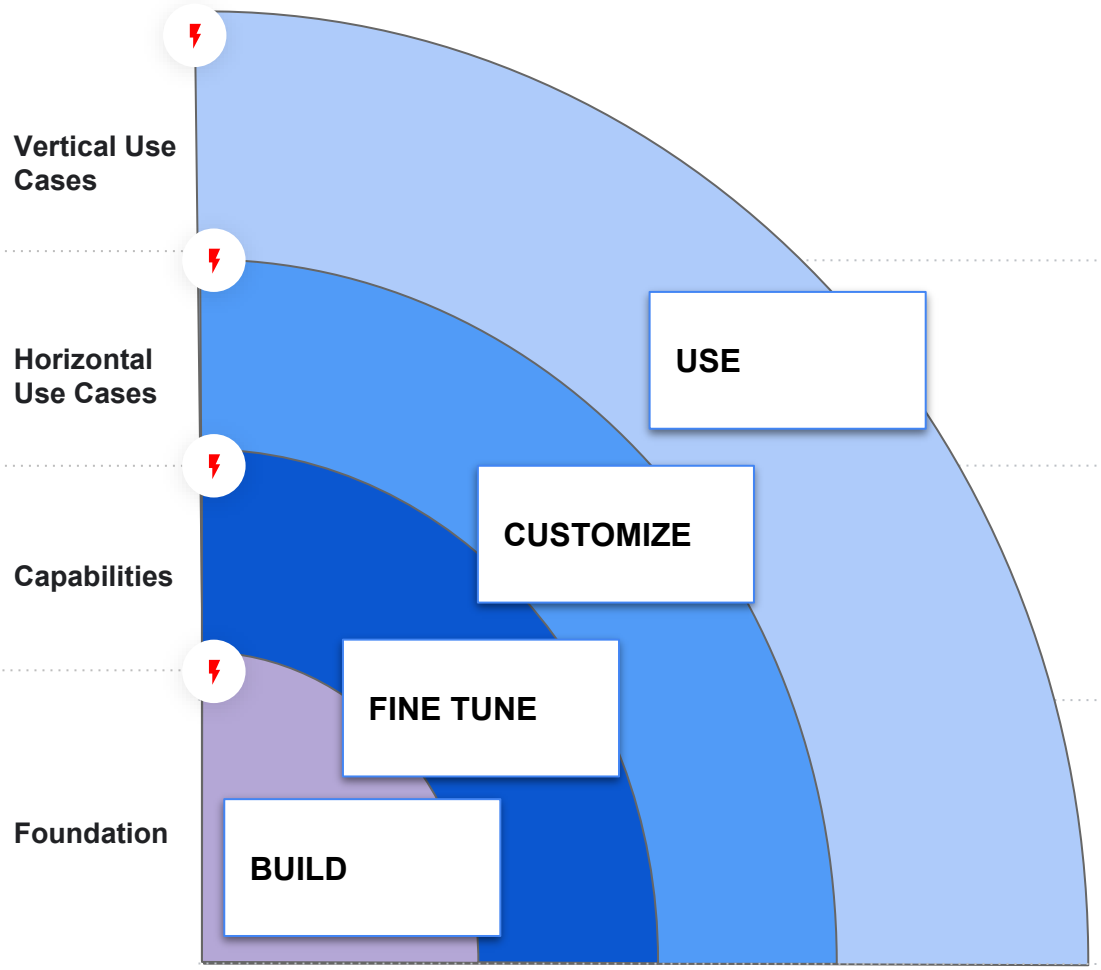
Experts Google Developers



The Layer Cake of GenAI

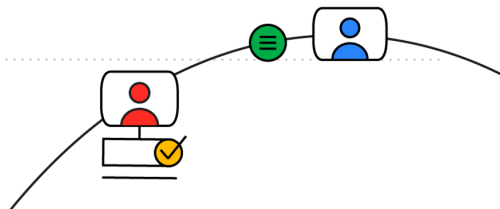
VINICARIDÁ
Human-Centered Data & AI

DevFest Cerrado 2024

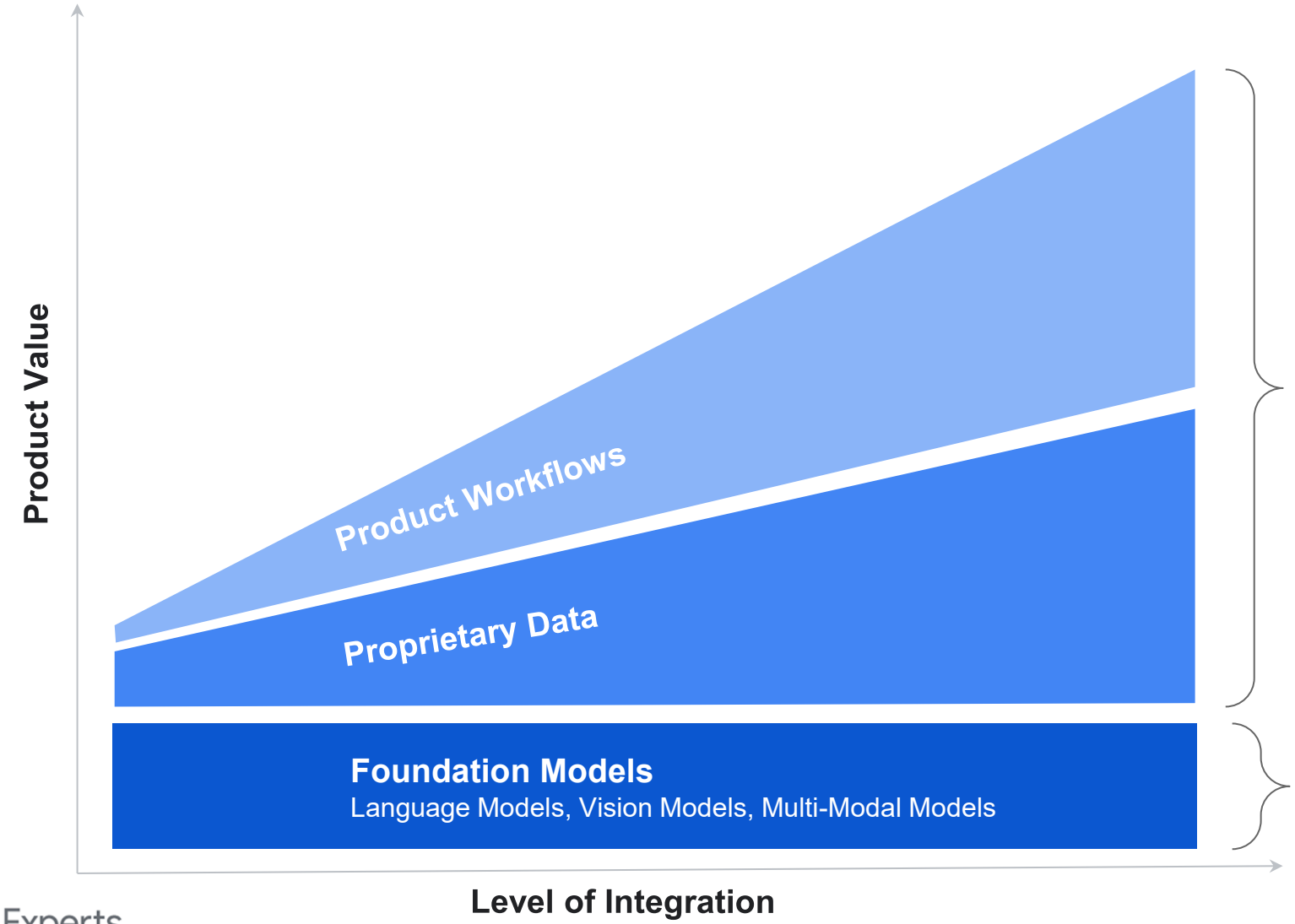


USE CASE	PERSONAS
COMPLIANCE ASSISTANT PERSONAL FINANCE ASSISTANT PERSONALIZED MKTG CONTENT KNOW YOUR CUSTOMER PROFILES FLEET MANAGEMENT	HR & FINANCE END USERS CUSTOMER RELATIONSHIP MANAGERS STRATEGIC OPERATIONS MARKETERS
SEARCH CONVERSATION PRODUCTIVITY ASK YOUR DATA	DATA ANALYSTS SOFTWARE ENGINEERS IT ANALYSTS
GENERATION SUMMARIZATION EXTRACTION DOCUMENT UNDERSTANDING MODAL TRANSITIONS	AI ENGINEERS PROMPT ENGINEERS SOFTWARE ENGINEERS
MODEL-OPS TRAINING PREDICTIVE/ANALYTICAL AI INFRASTRUCTURE DATASETS	MODEL ENGINEERS DATA ENGINEERS DATA SCIENTISTS INFRA ENGINEERS

Experts
Google Developers

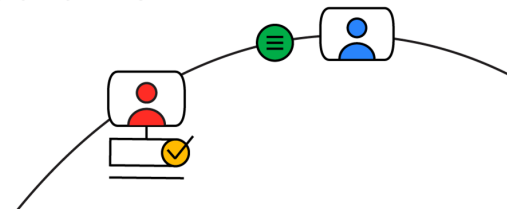


Where Value Accrues with GenAI



Your Competitive Advantage

Everyone has access to this

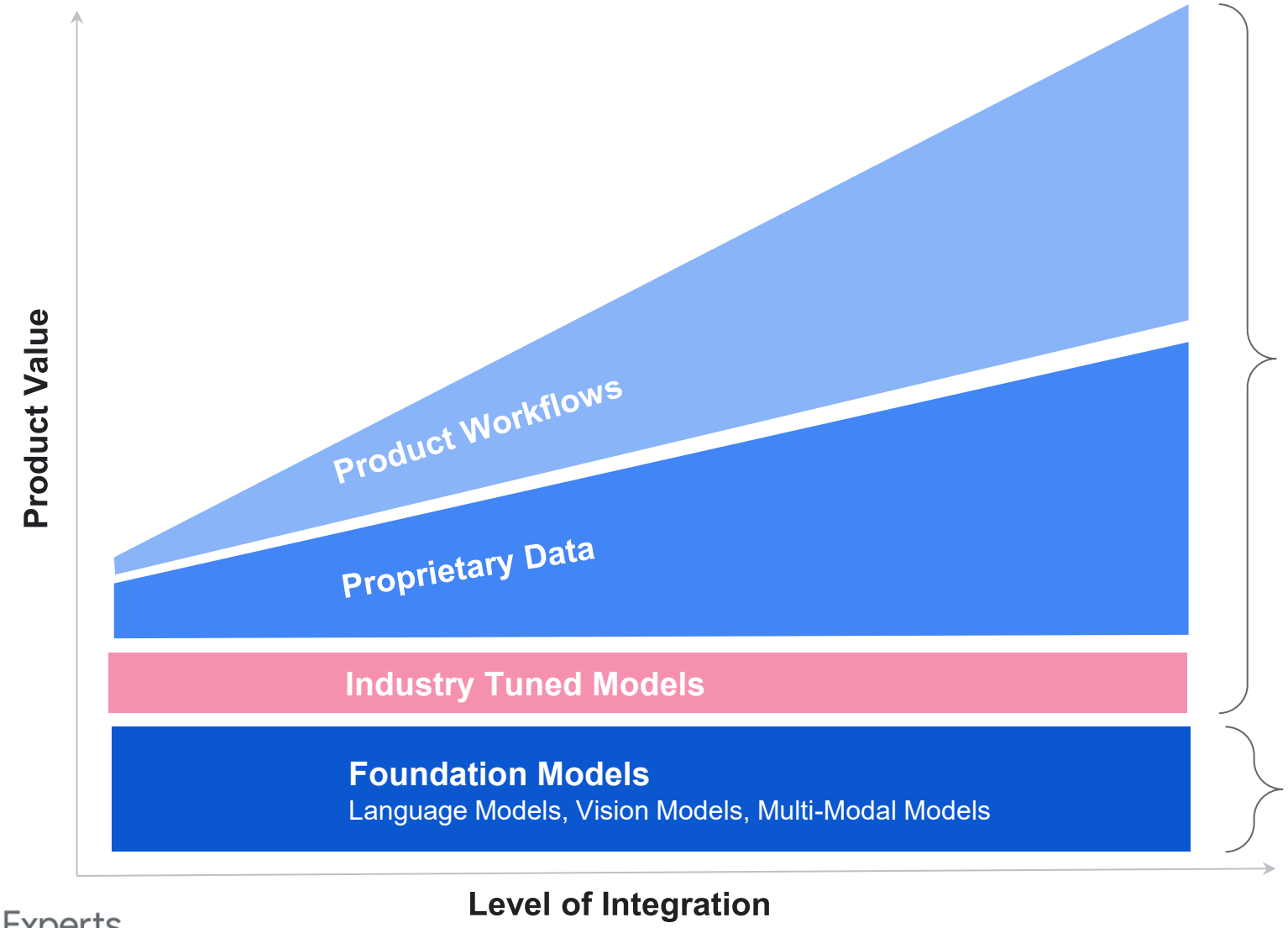


VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Cerrado

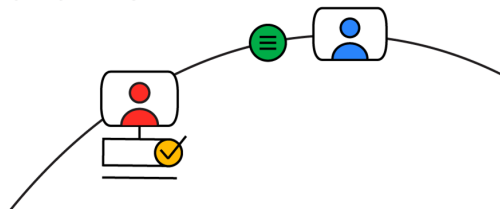
Experts
Google Developers

Where Value Accrues with GenAI



Your Competitive Advantage

Everyone has access to this



VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Cerrado

Experts
Google Developers

VINICARIDÁ
Human-Centered Data & AI

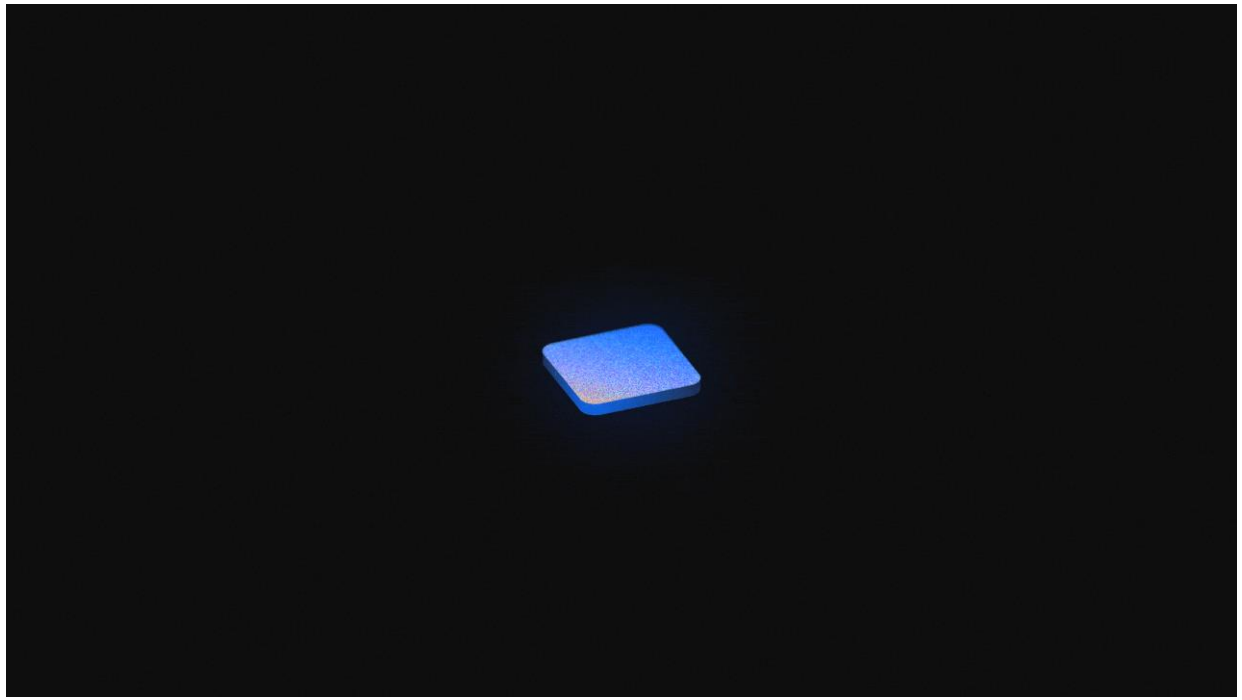
04 The Gemini and Gemma Era

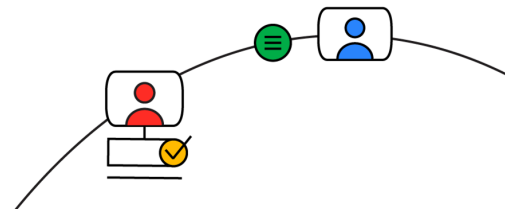
DevFest 2024 Google Developer Groups Cerrado

The next chapter of Generative AI innovation

Gemini

Gemini is the most capable and general model we've ever built, and is the result of a large-scale collaborative effort by teams across Google, including Google DeepMind and Google Research.

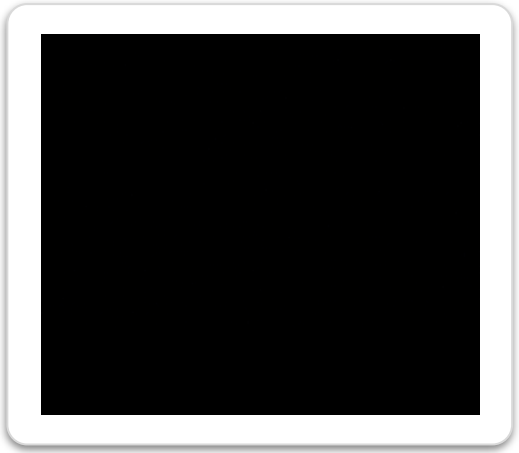




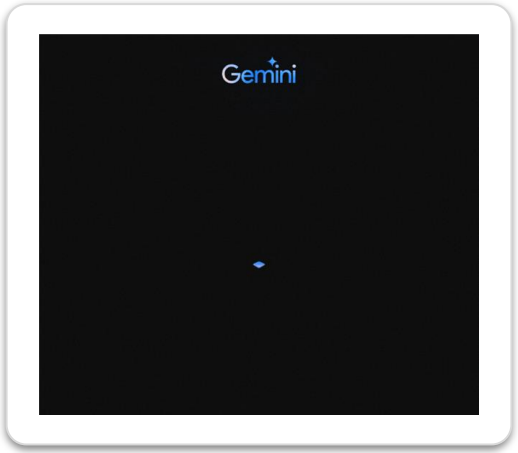
VINICARIDÁ
Human-Centered Data & AI

Gemini marks the next phase on our journey to making AI more helpful for everyone

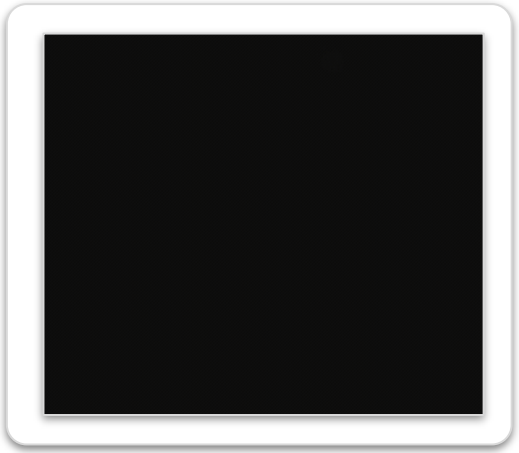
DevFest 2024 Cerrado



State-of-the-art, natively multimodal reasoning capabilities

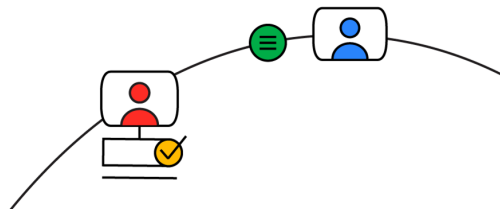


Highly optimized while preserving choice



Built with responsibility and safety at the core

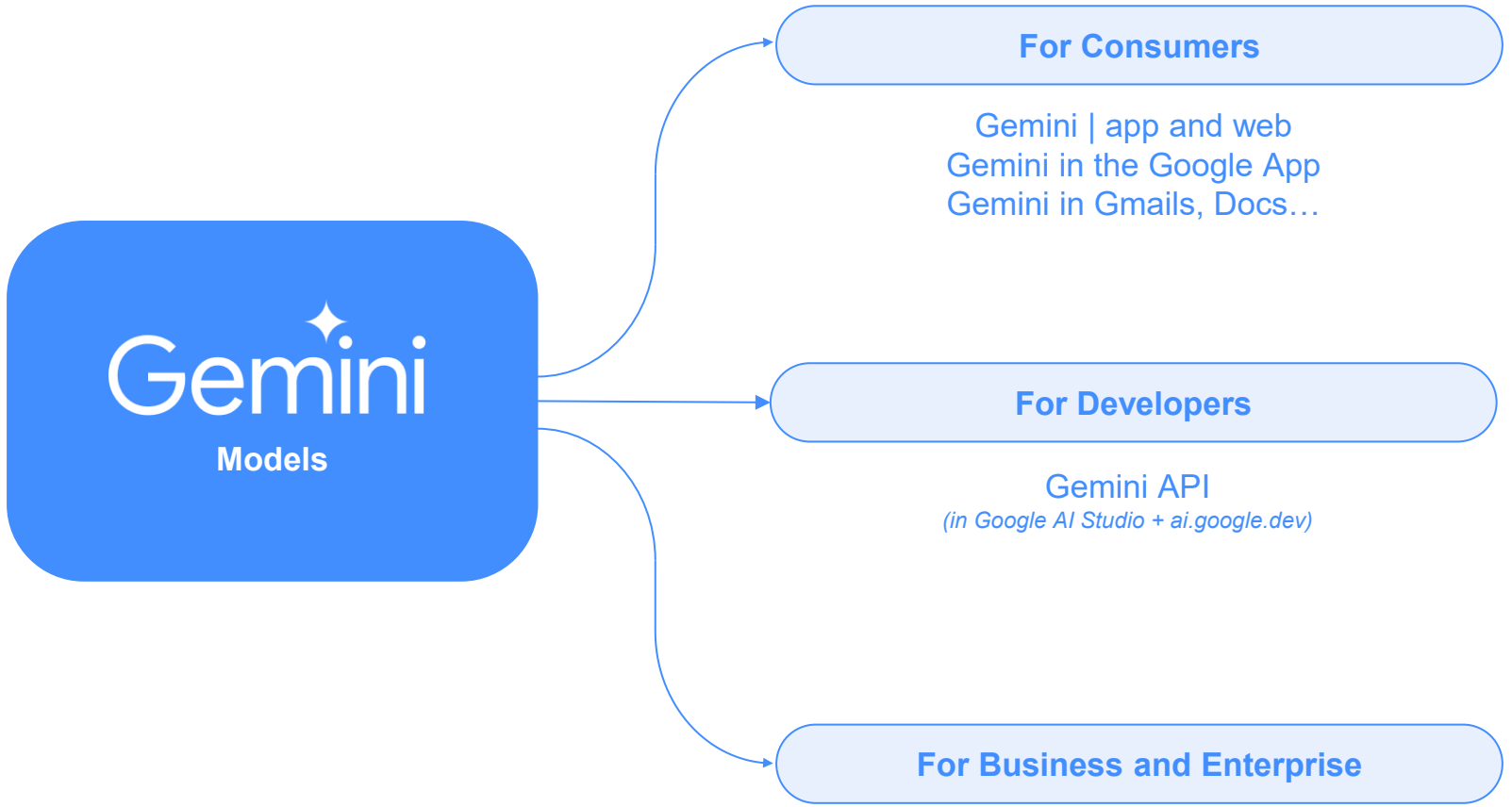
Experts
Google Developers



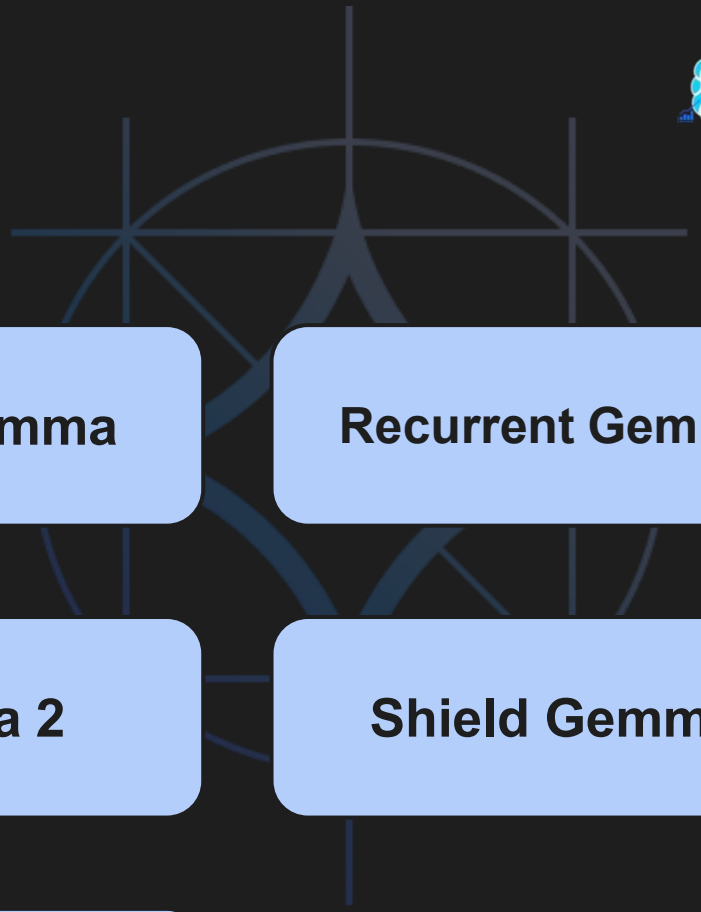
The Gemini Ecosystem

The most advanced AI from Google

VINICARIDÁ
Human-Centered Data & AI



Gemma Family



VINICARIDÁ
Human-Centered Data & AI

- Gemma
- Code Gemma
- Recurrent Gemma
- Pali Gemma
- Gemma 2
- Shield Gemma
- Gemma Scope

DevFest 2024

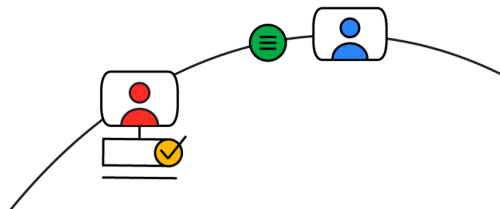
VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Google Developer Groups Cerrado

kaggle

Gemma

Experts
Google Developers



kaggle Competitions Datasets Models Code Discussions Courses ... Search Sign In Register

Level up with the largest AI & ML community

Join over 17M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

Register with Google Register with Email

Who's on Kaggle?

Learners
Dive into Kaggle courses, competitions & forums.

KEY FEATURES

- Beginner competitions
- Practical courses
- Public datasets

Developers
Leverage Kaggle's models, notebooks & datasets.

KEY FEATURES

- Open-sourced models
- Competition solution write-ups
- Public notebooks

Researchers
Advance ML with our pre-trained model hub & competitions.

KEY FEATURES

- ML competition hosting
- Model training datasets
- Open-sourced models

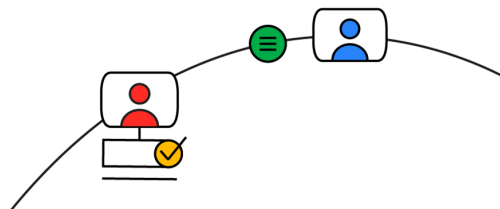
VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Google Developer Groups Cerrado

Keras

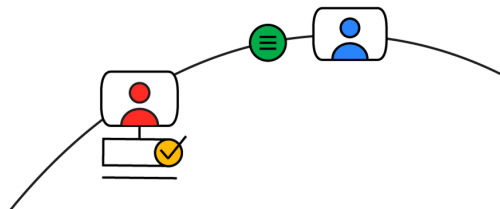
Gemma

Experts
Google Developers



Why Keras?

Simple	Flexible	Powerful
simplifies development, empowering focused problem-solving.	offers scalable complexity for evolving needs.	industry-strength performance and scalability



Vertex AI

Build your own generative AI-powered agents

AI Solutions
Contact Center AI | Document AI | Risk AI | …

Search

Conversation

AI Platform
Extensions | Connectors | Grounding
Prompt | Serve | Tune | Distill | Eval

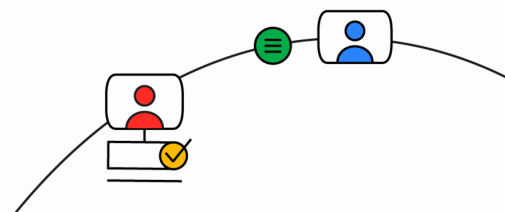
Gemini and 130+ models

Model Garden
Google | OSS | Partner Models

Gemma

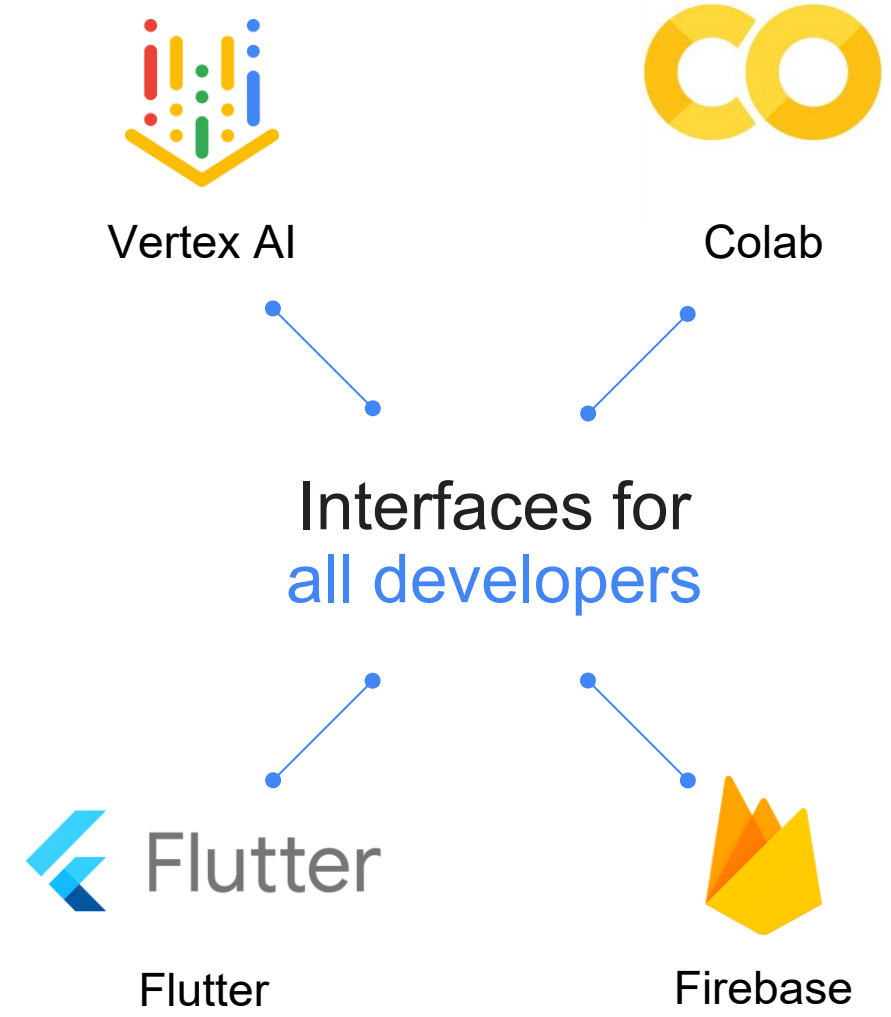
Google Cloud Infrastructure (GPU/TPU) | Google Data Cloud

- Business Users
- Developers
- AI Practitioners



Vertex AI is built for developers

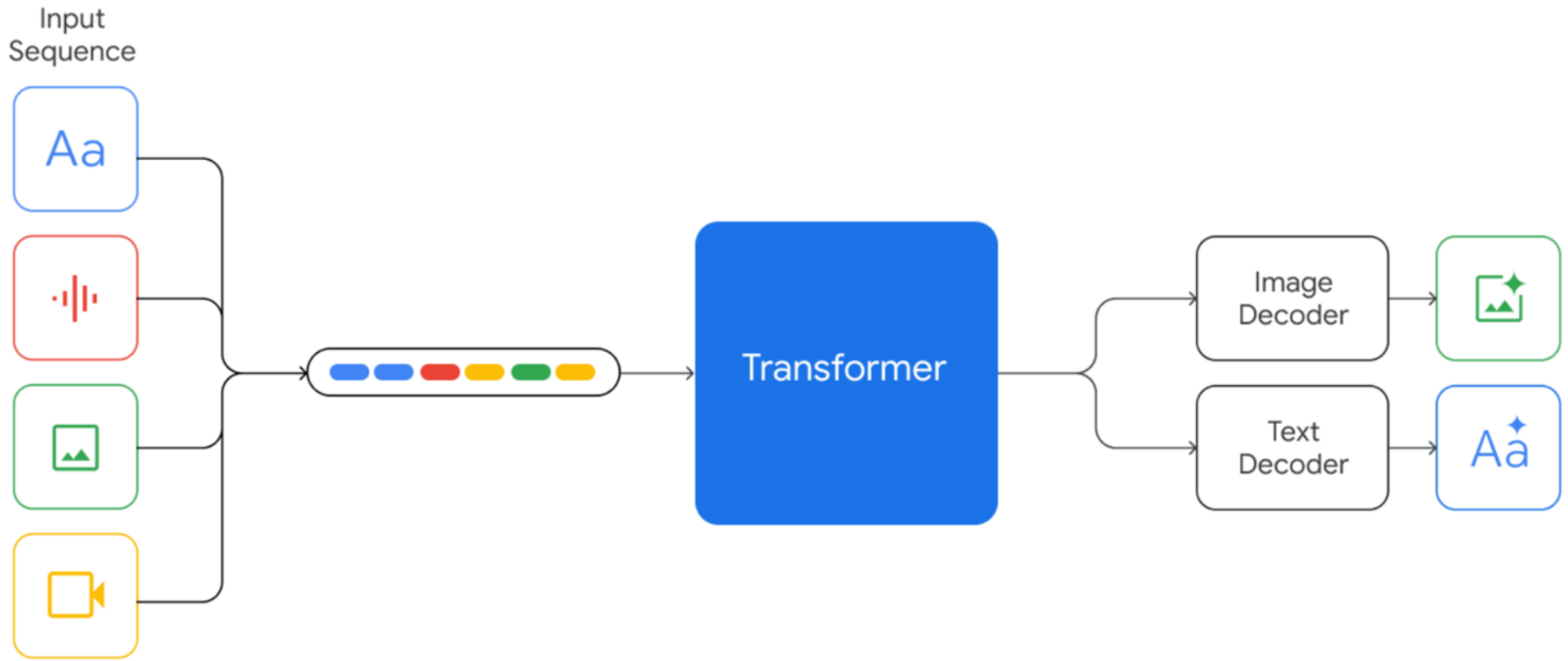
- Extensive **quick start library** with code samples and jumpstarts for **developers of all levels** and ecosystems
- **Free developer labs** and training resources across Vertex products at Cloud Skills Boost
- **Robust integrations** with popular third party developer tools like **Lang Chain, LlamaIndex, Pinecone, and Weaviate.**
- **Packages and extensions** to natively support Google Cloud foundation models in Google app developer frameworks like **Firebase and Flutter**.



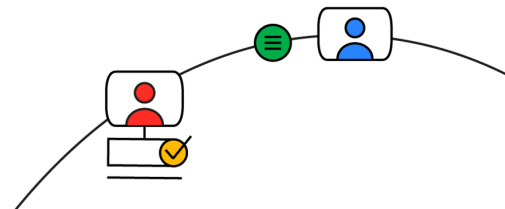
VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Cerrado

Multimodality



Experts
Google Developers

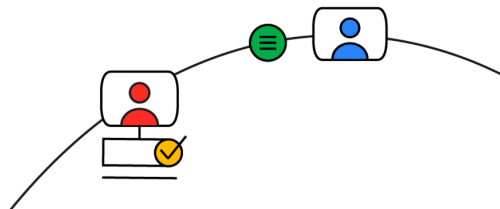


VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Google Developer Groups Cerrado

Gemini VS Gemma

Experts
Google Developers



Gemini

Gemma

VINICARIDÁ
Human-Centered Data & AI

DevFest 2024 Cerrado

Purpose

Gemini is Google's next-generation large language model (LLM). It's a more advanced model designed to compete with other state-of-the-art models in natural language understanding and generation, such as GPT-4. Gemini is built to excel in tasks that require deep understanding, reasoning, and creativity, specifically in conversational AI and content generation.

Gemma is designed to be a highly versatile AI model, mainly focused on general-purpose machine learning tasks like natural language processing (NLP), computer vision, and other structured or unstructured data tasks.

Experts
Google Developers

Gemini

Gemma

VINICARIDÁ
Human-Centered Data & AI

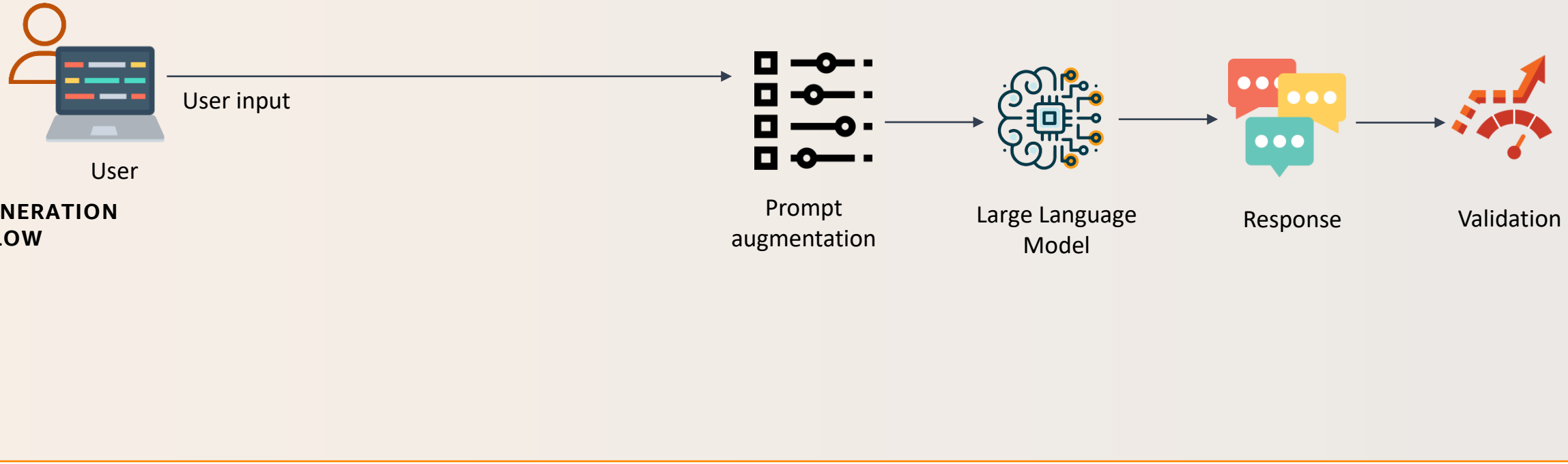
DevFest 2024 Cerrado

Integration

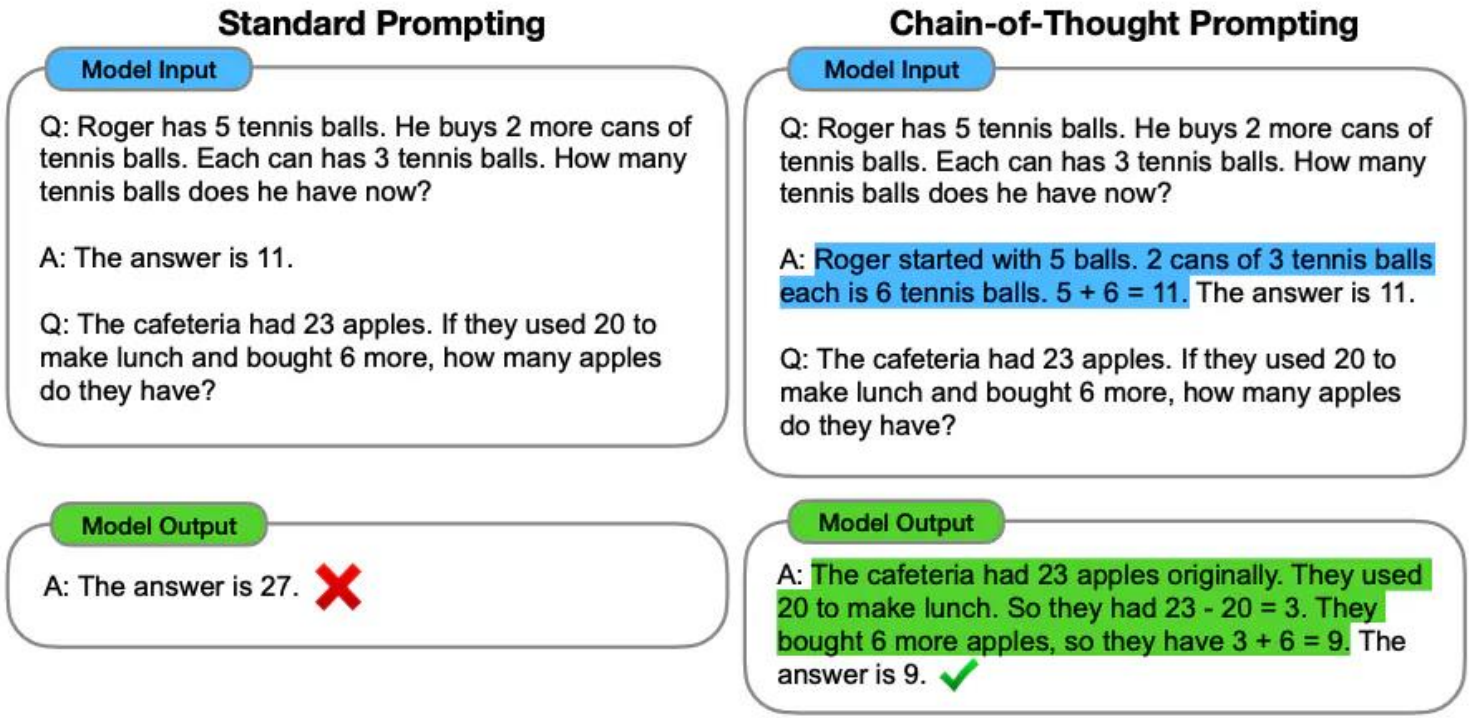
Gemini is integrated into Google services like Bard and Search, and can be used via Google Cloud for NLP-heavy applications.

It's also deeply integrated into Google Cloud's Vertex AI ecosystem, making it easy to use for developers and data scientists who are already working within the Google Cloud environment.

Experts
Google Developers

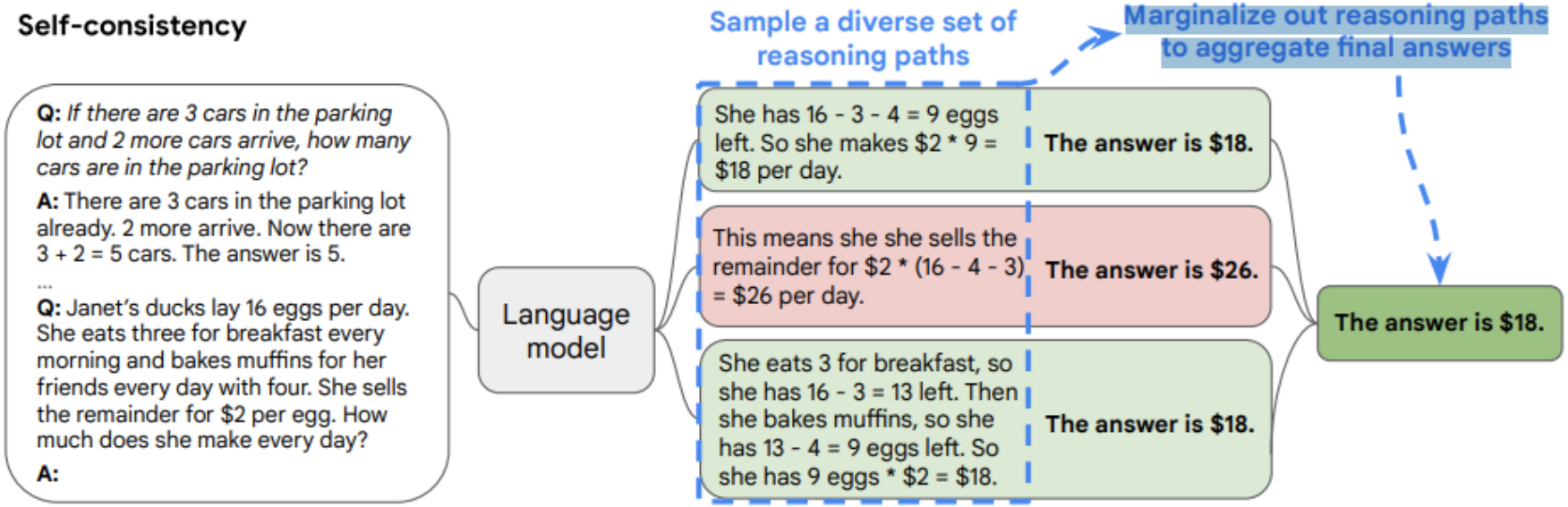


Chain-of-Thought Prompting (CoT)

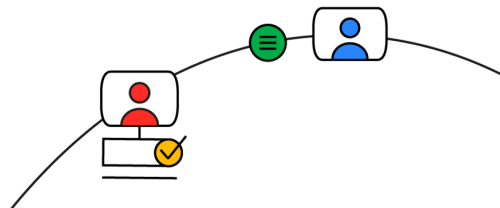


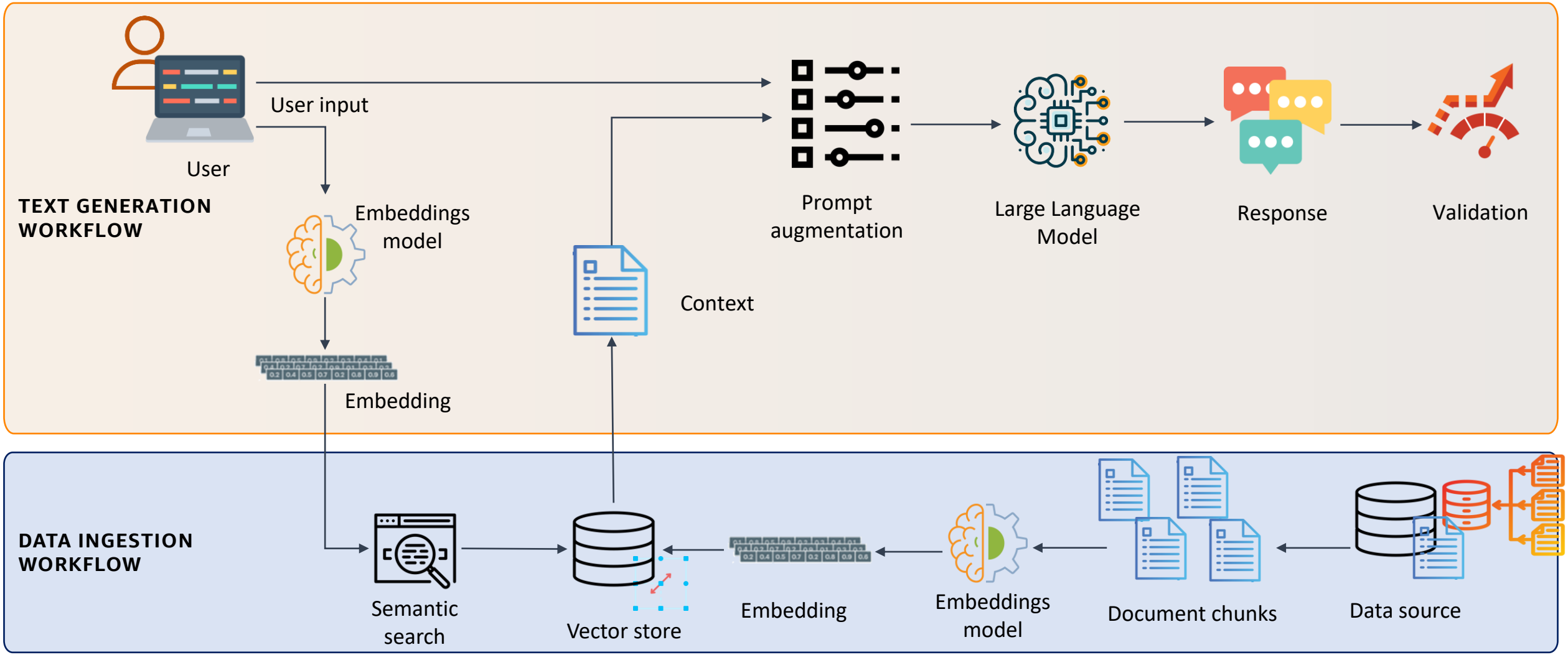
Ref: https://arxiv.org/abs/2201.11903

Self Consistency



Ref: https://arxiv.org/pdf/2203.11171



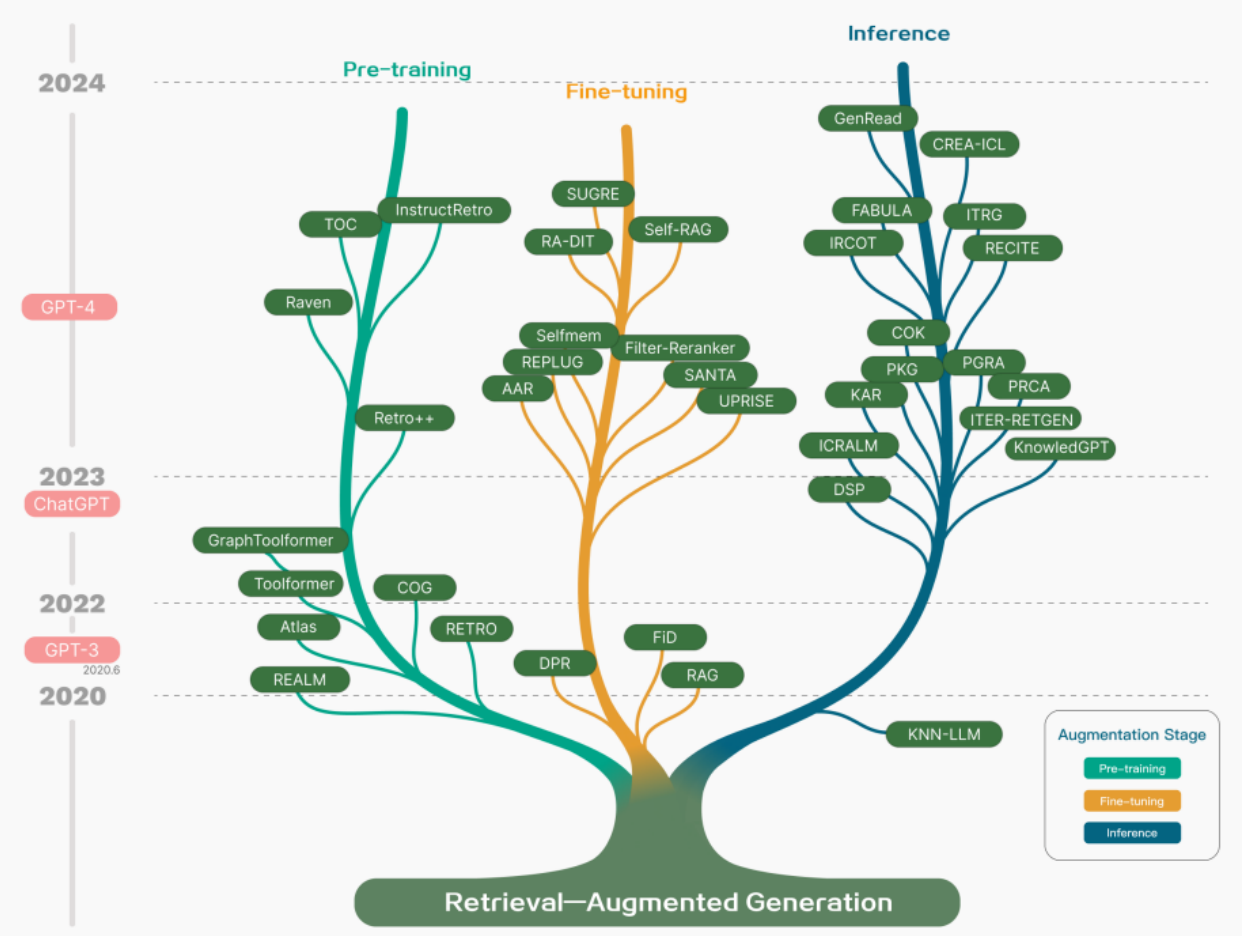


VINICARIDÁ
Human-Centered Data & AI

Retrieval-Augmented Generation for Large Language Models: A Survey

Yunfan Gao [1], **Yun Xiong** [2], **Xinyu Gao** [2], **Kangxiang Jia** [2], **Jinliu Pan** [2], **Yuxi Bi** [3], **Yi Dai** [1], **Jiawei Sun** [1], **Qianyu Guo** [4], **Meng Wang** [3] and **Haofen Wang** [1,3] *

[1] Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University
[2] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[3] College of Design and Innovation, Tongji University
[4] School of Computer Science, Fudan University



https://arxiv.org/pdf/2312.10997.pdf

THE CHRONICLES OF RAG: THE RETRIEVER, THE CHUNK AND THE GENERATOR

PREPRINT

*Paulo Finardi *Leonardo Avila Rodrigo Castaldoni Pedro Gengo Celio Larcher
Marcos Piau Pablo Costa Vinicius Caridá

22h, Brazil

email: {pfinardi, leonardo.bernardi.avila, castaldoniro, pedro.gengo.lourenco, celiolarcher, marcos.piau.vieira, pablo.botton.costa, vfcarida}@gmail.com

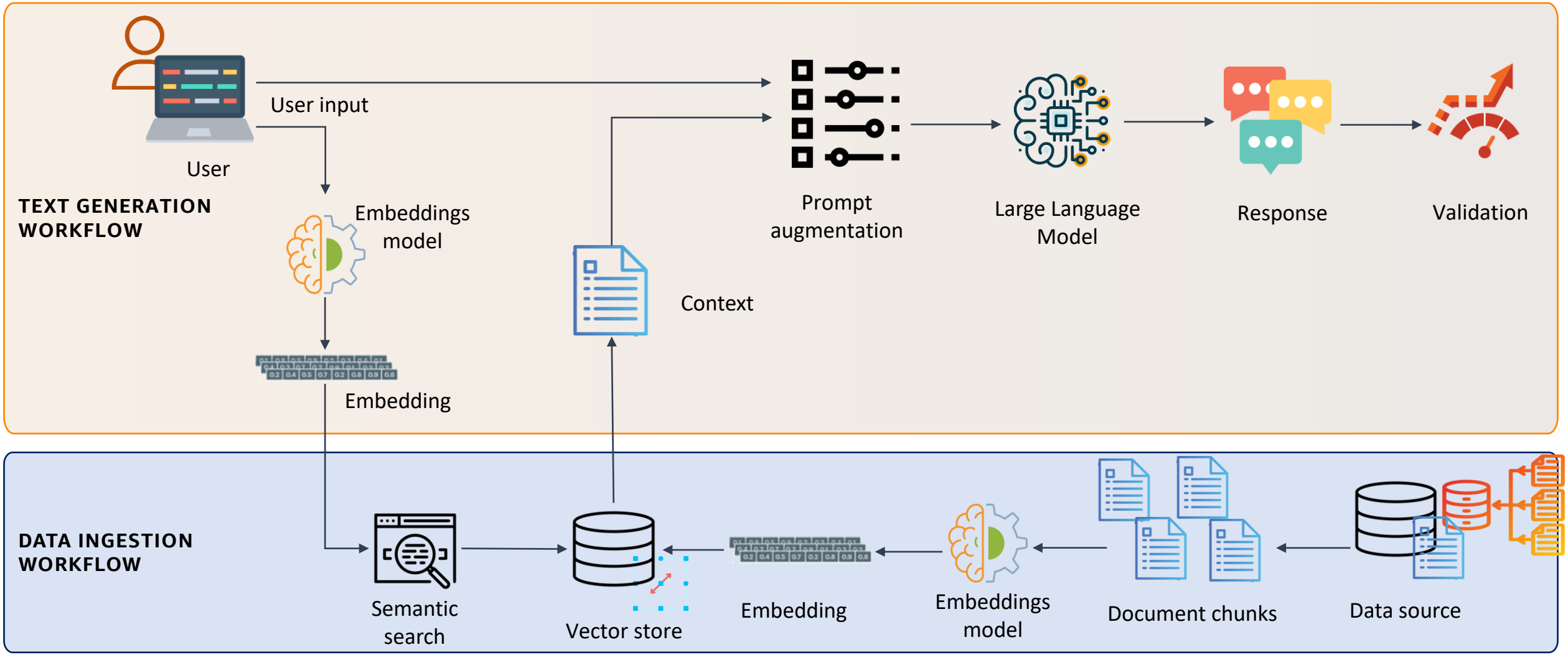
* Both authors contributed equally to this research.

ABSTRACT

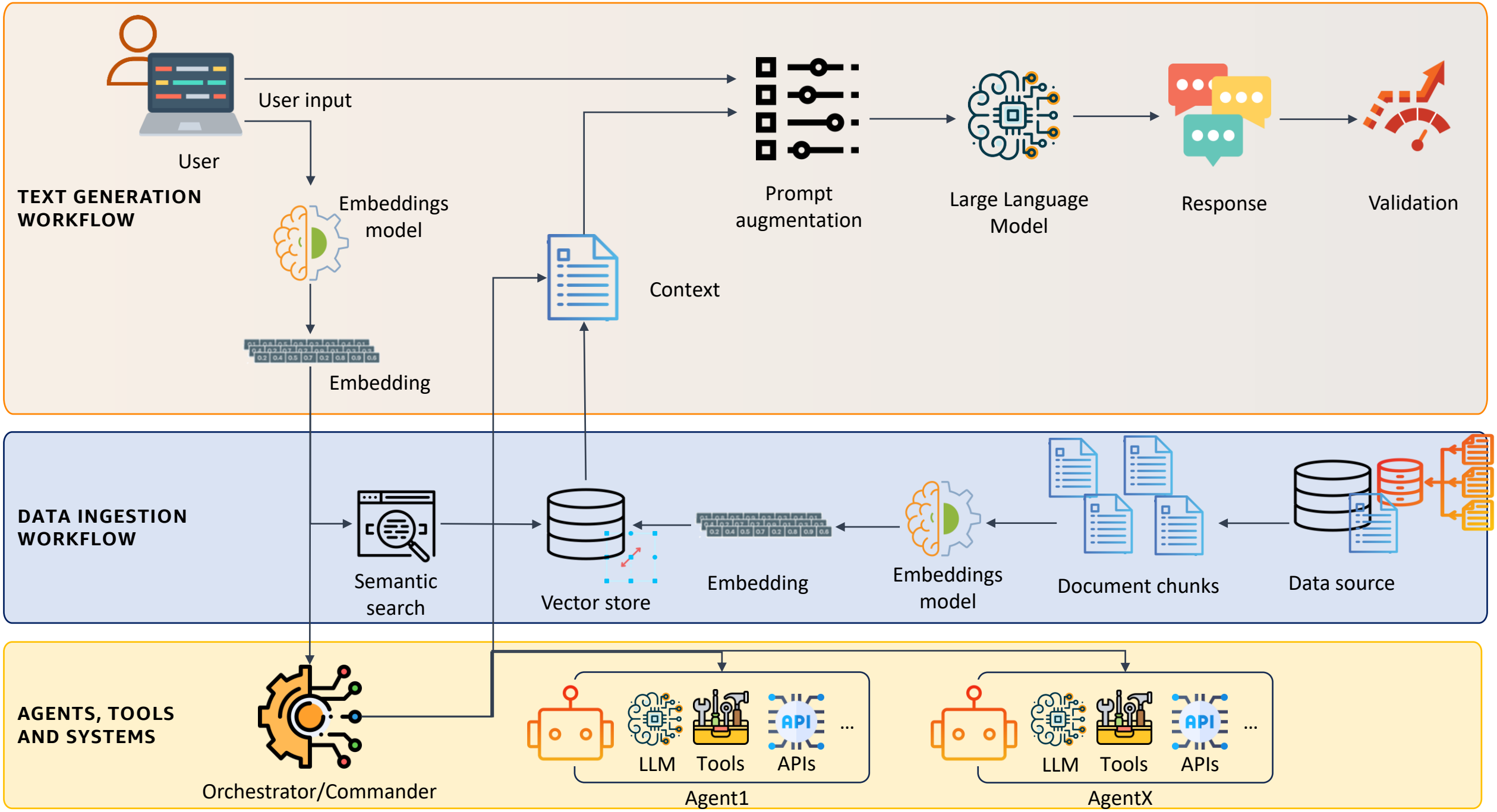
Retrieval Augmented Generation (RAG) has become one of the most popular paradigms for enabling LLMs to access external data, and also as a mechanism for grounding to mitigate against hallucinations. When implementing RAG you can face several challenges like effective integration of retrieval models, efficient representation learning, data diversity, computational efficiency optimization, evaluation, and quality of text generation. Given all these challenges, every day a new technique to improve RAG appears, making it unfeasible to experiment with all combinations for your problem. In this context, this paper presents good practices to implement, optimize, and evaluate RAG for the Brazilian Portuguese language, focusing on the establishment of a simple pipeline for inference and experiments. We explored a diverse set of methods to answer questions about the first Harry Potter book. To generate the answers we used the OpenAI's `gpt-4`, `gpt-4-1106-preview`, `gpt-3.5-turbo-1106`, and Google's `Gemini Pro`. Focusing on the quality of the retriever, our approach achieved an improvement of MRR@10 by 35.4% compared to the baseline. When optimizing the input size in the application, we observed that it is possible to further enhance it by 2.4%. Finally, we present the complete architecture of the RAG with our recommendations. As result, we moved from a baseline of 57.88% to a maximum relative score of 98.61%.

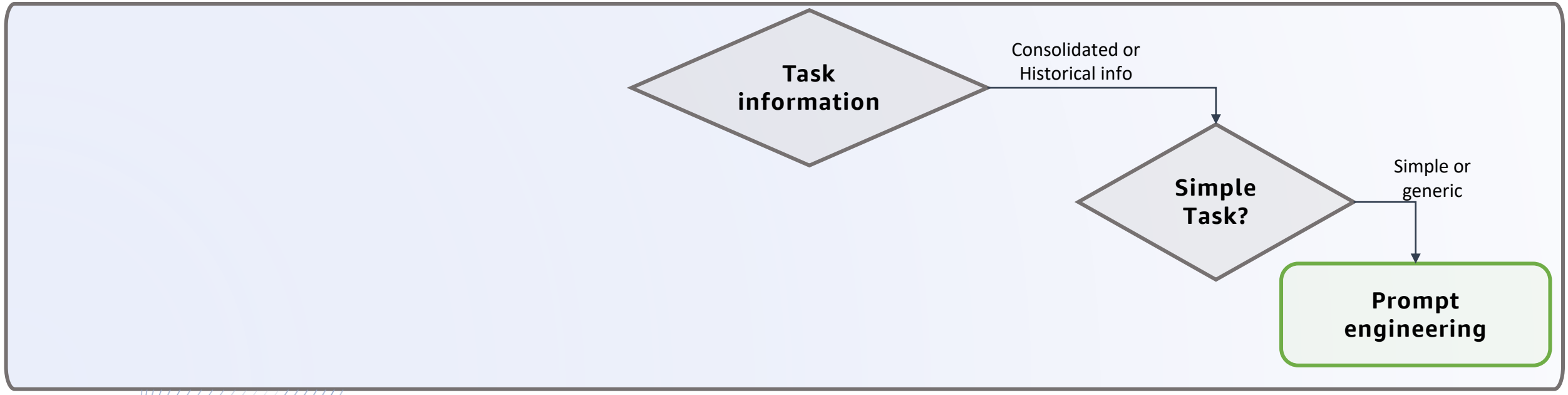
https://arxiv.org/abs/2401.07883

VINICARIDÁ
Human-Centered Data & AI



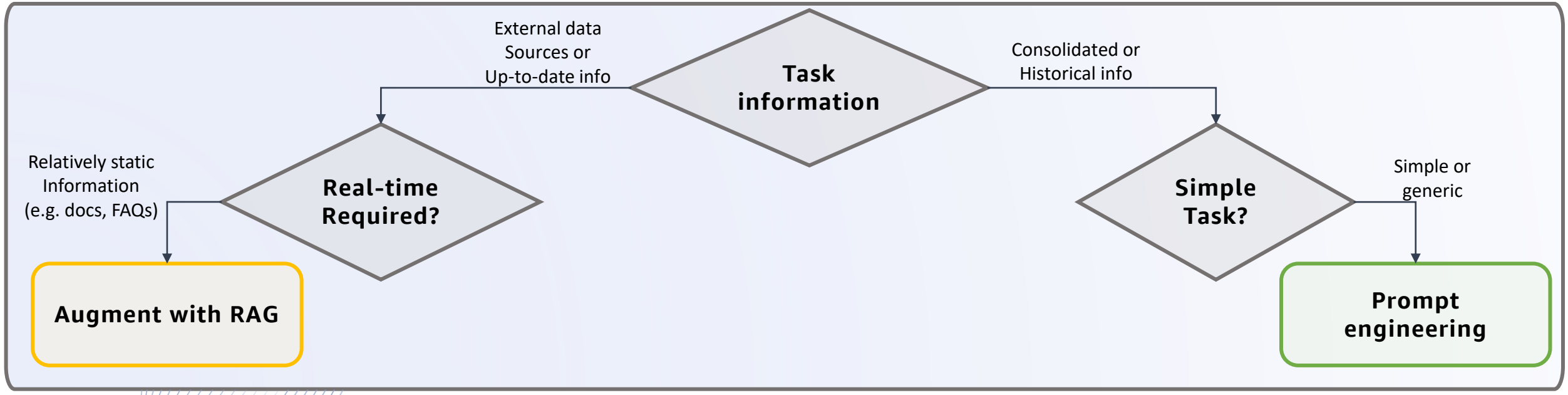
VINICARIDÁ
Human-Centered Data & AI


User input
User
TEXT GENERATION WORKFLOW
Embeddings model
Embedding
Context
Prompt augmentation
Large Language Model
Response
Validation
DATA INGESTION WORKFLOW
Semantic search
Vector store
Embedding
Embeddings model
Document chunks
Data source
AGENTS, TOOLS AND SYSTEMS
Orchestrator/Commander
LLM
Tools
APIs
...
Agent1
AgentX



Prompt Engineering

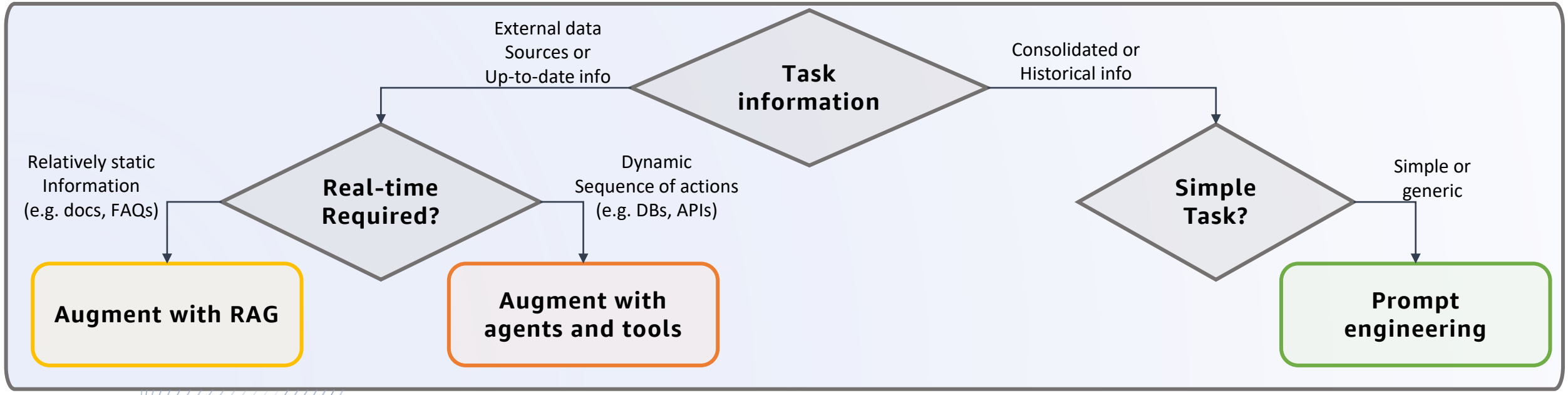
VINI CARIDÁ
Human-Centered Data & AI



Prompt Engineering

Retrieval Augmented Generation (RAG)

VINI CARIDÁ
Human-Centered Data & AI

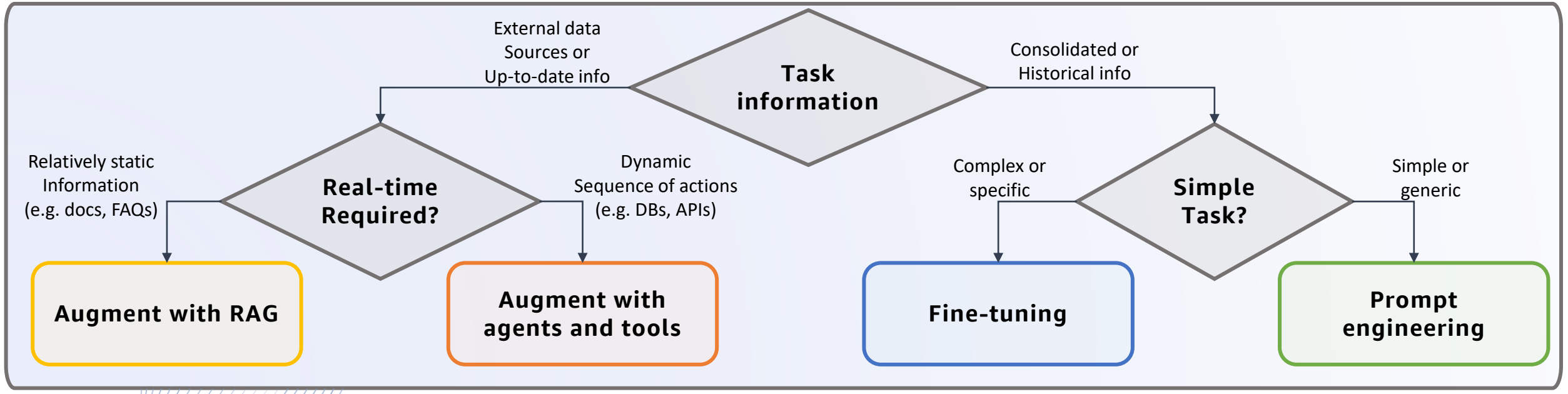


Prompt Engineering

Retrieval Augmented Generation (RAG)

Agents and Tools

VINI CARIDÁ
Human-Centered Data & AI



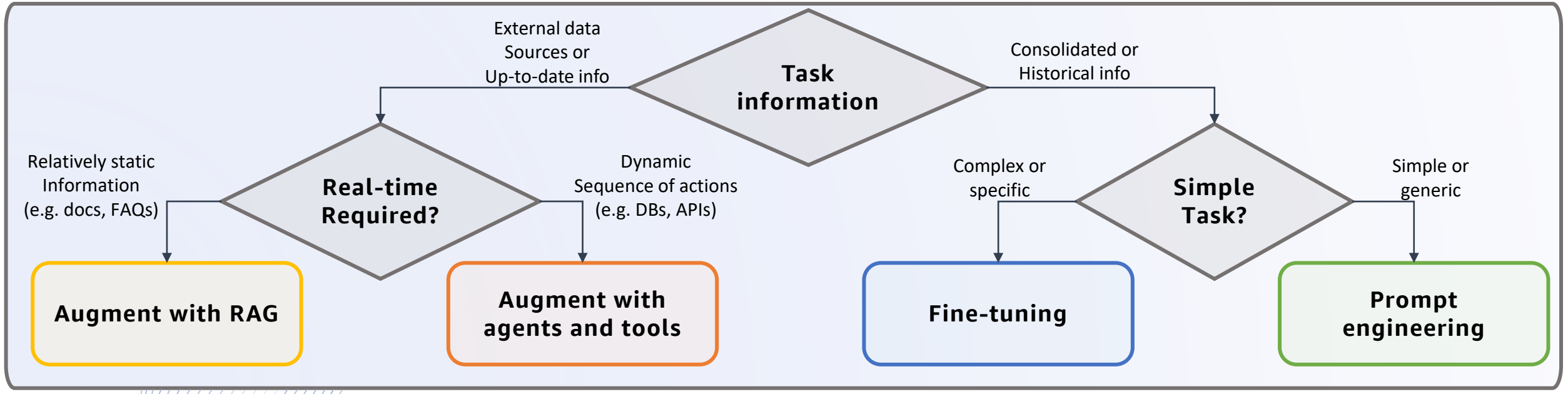
Prompt Engineering

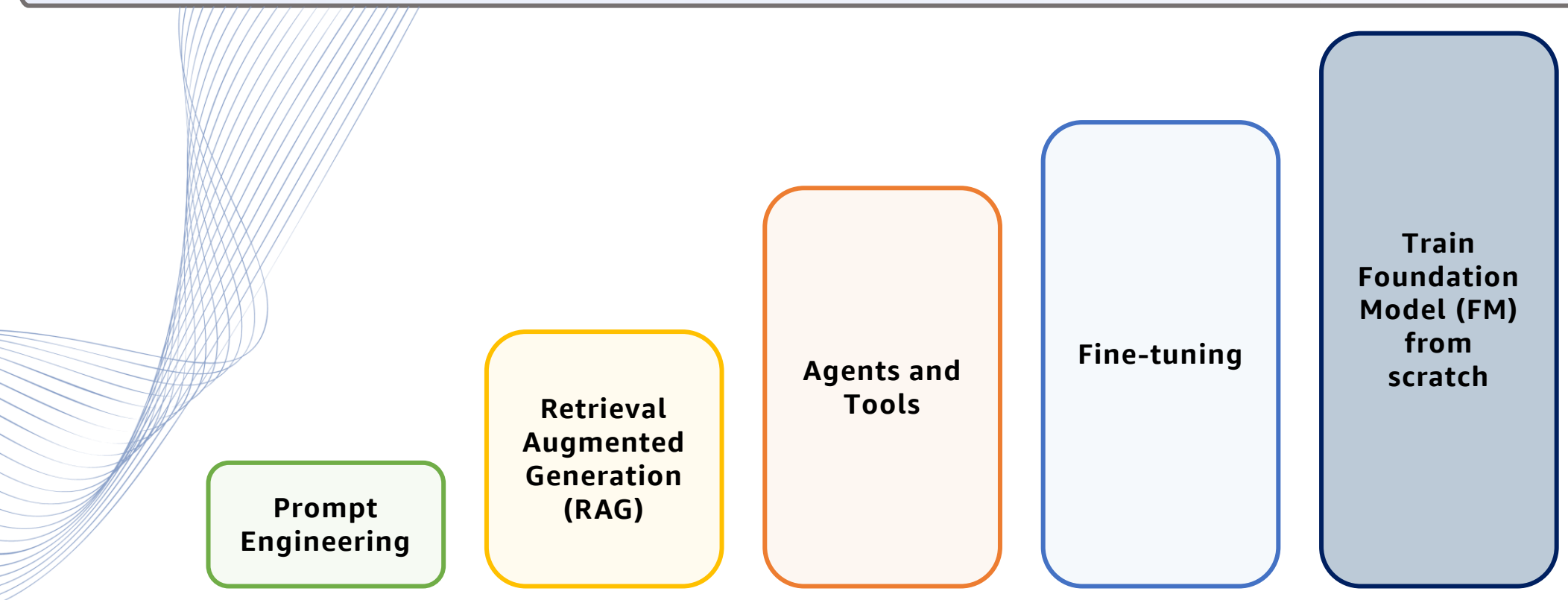
Retrieval Augmented Generation (RAG)

Agents and Tools

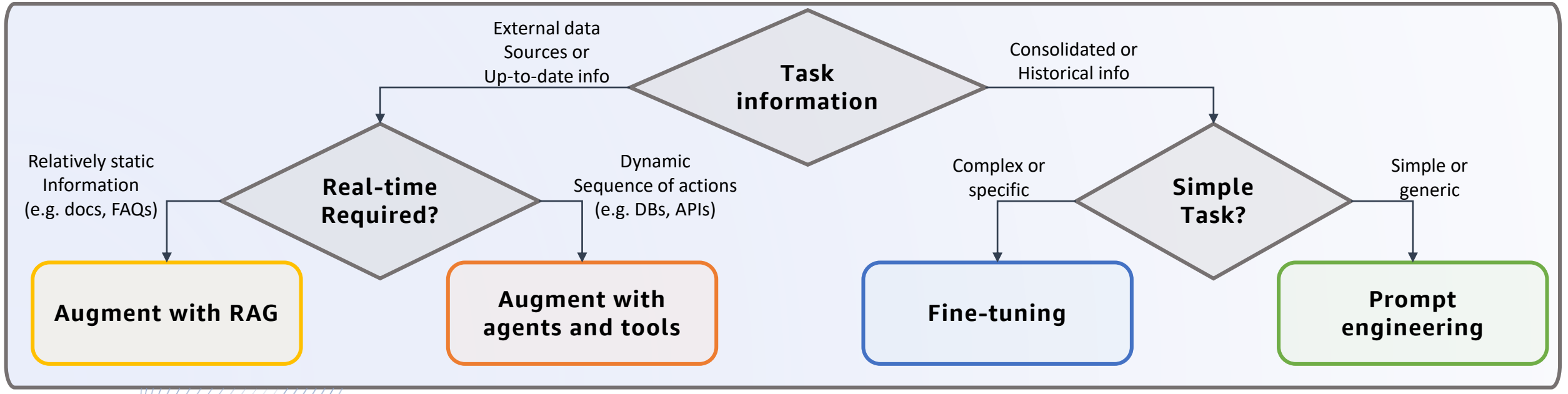
Fine-tuning

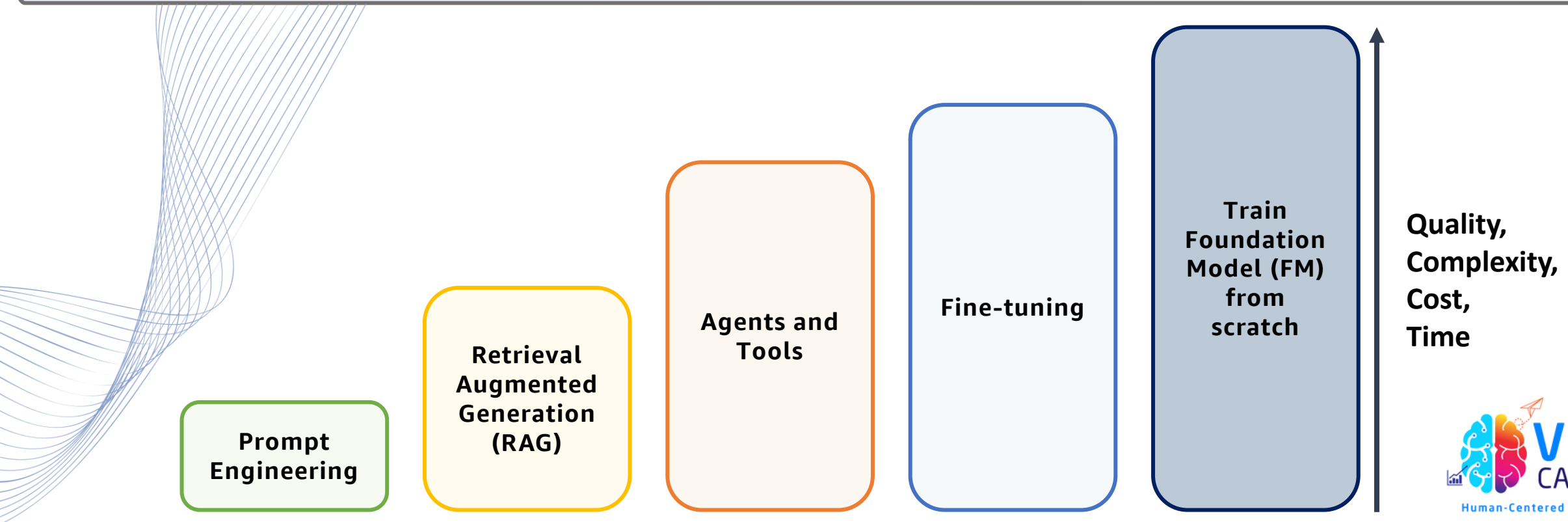
VINI CARIDÁ
Human-Centered Data & AI





VINI CARIDÁ
Human-Centered Data & AI





VINI CARIDÁ
Human-Centered Data & AI

VINICARIDÁ
Human-Centered Data & AI

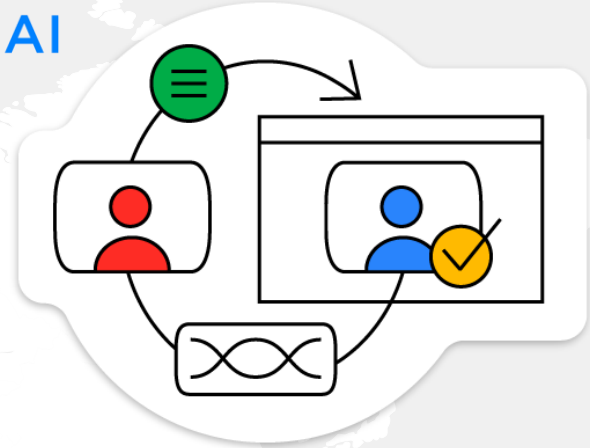
DevFest 2024 Cerrado

Bora ver na prática!

Experts
Google Developers

VINICARIDÁ
Human-Centered Data & AI

Thank you!



@vinicius caridá

@vfcarida

vfcarida@gmail.com

https://linktr.ee/vfcarida