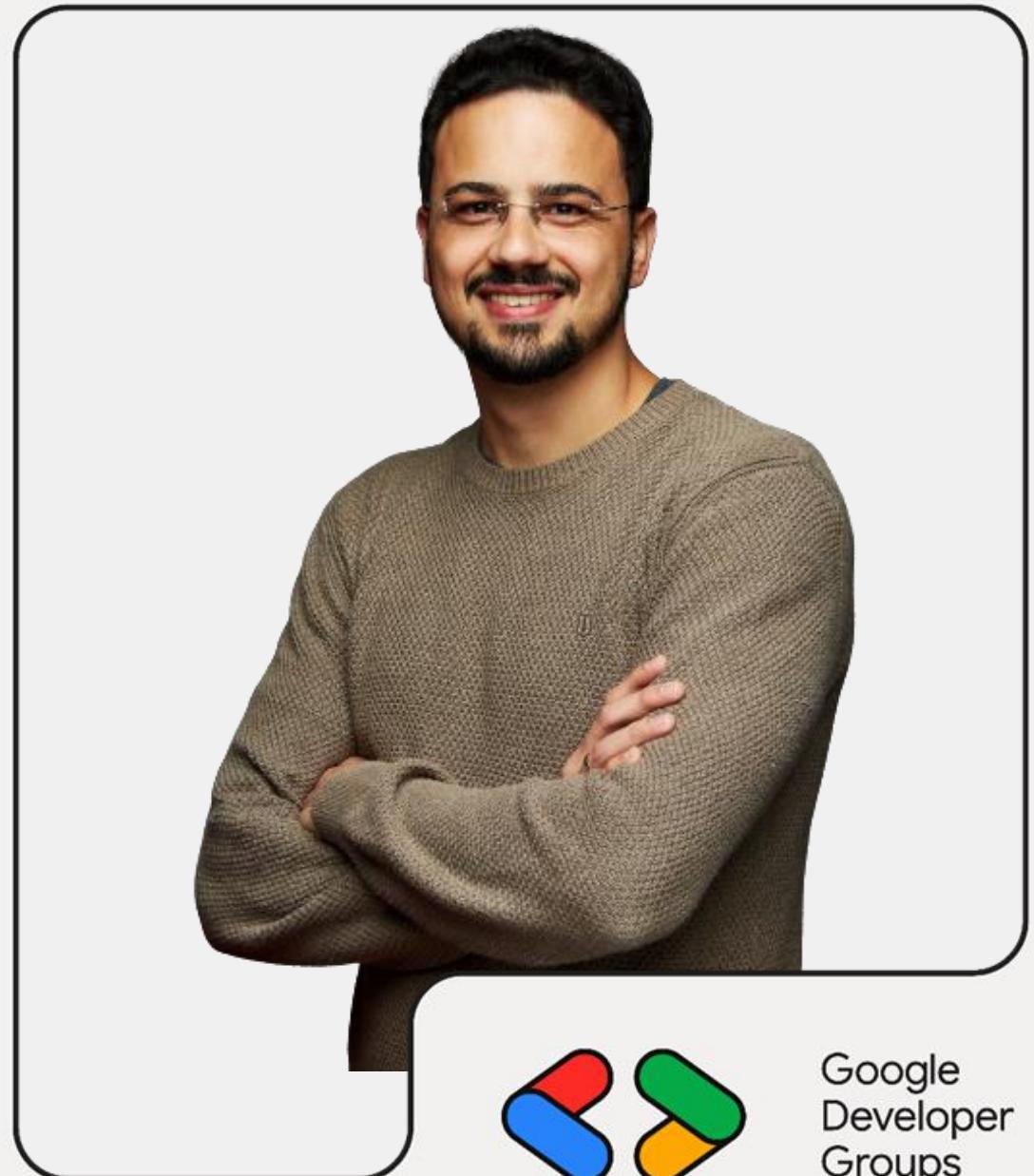
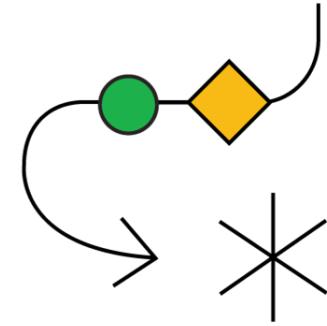
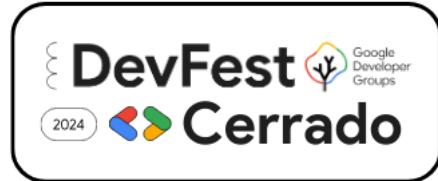


# Vini Caridá

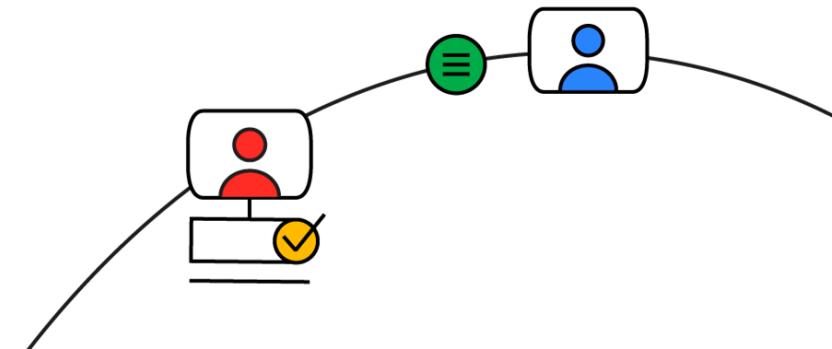
Usando Tecnologia, Dados e AI para alavancar negócios e impactar pessoas para um mundo mais justo e evoluido.



Google  
Developer  
Groups



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





# 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

01

# GenAI, Hype ou Realidade?

# Generative AI

Artificial Intelligence

Machine Learning

Deep Learning

Generative 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

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

# Generative AI | Hype cycle for AI

2022



[gartner.com](http://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

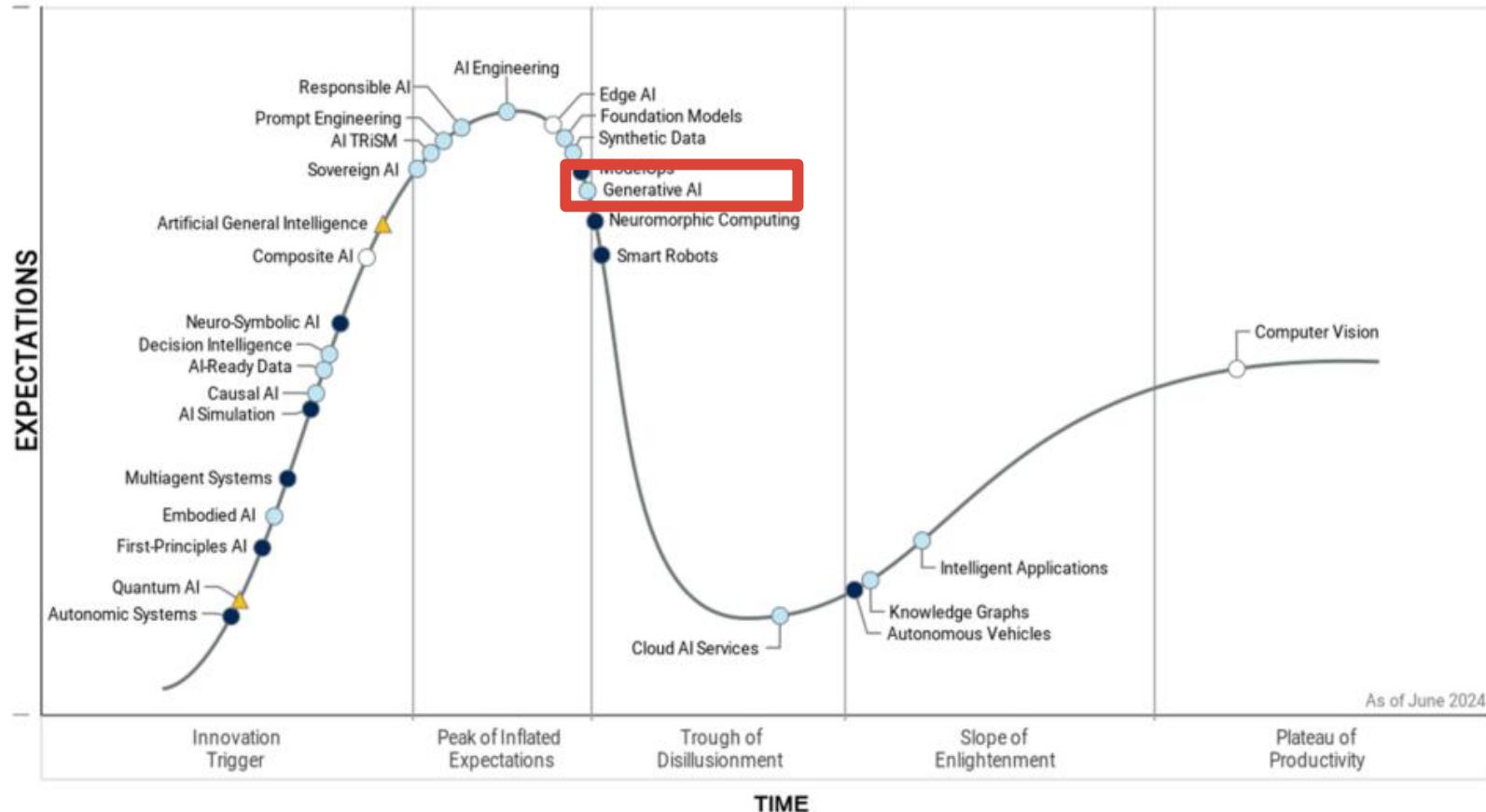


[gartner.com](http://gartner.com)

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

2024

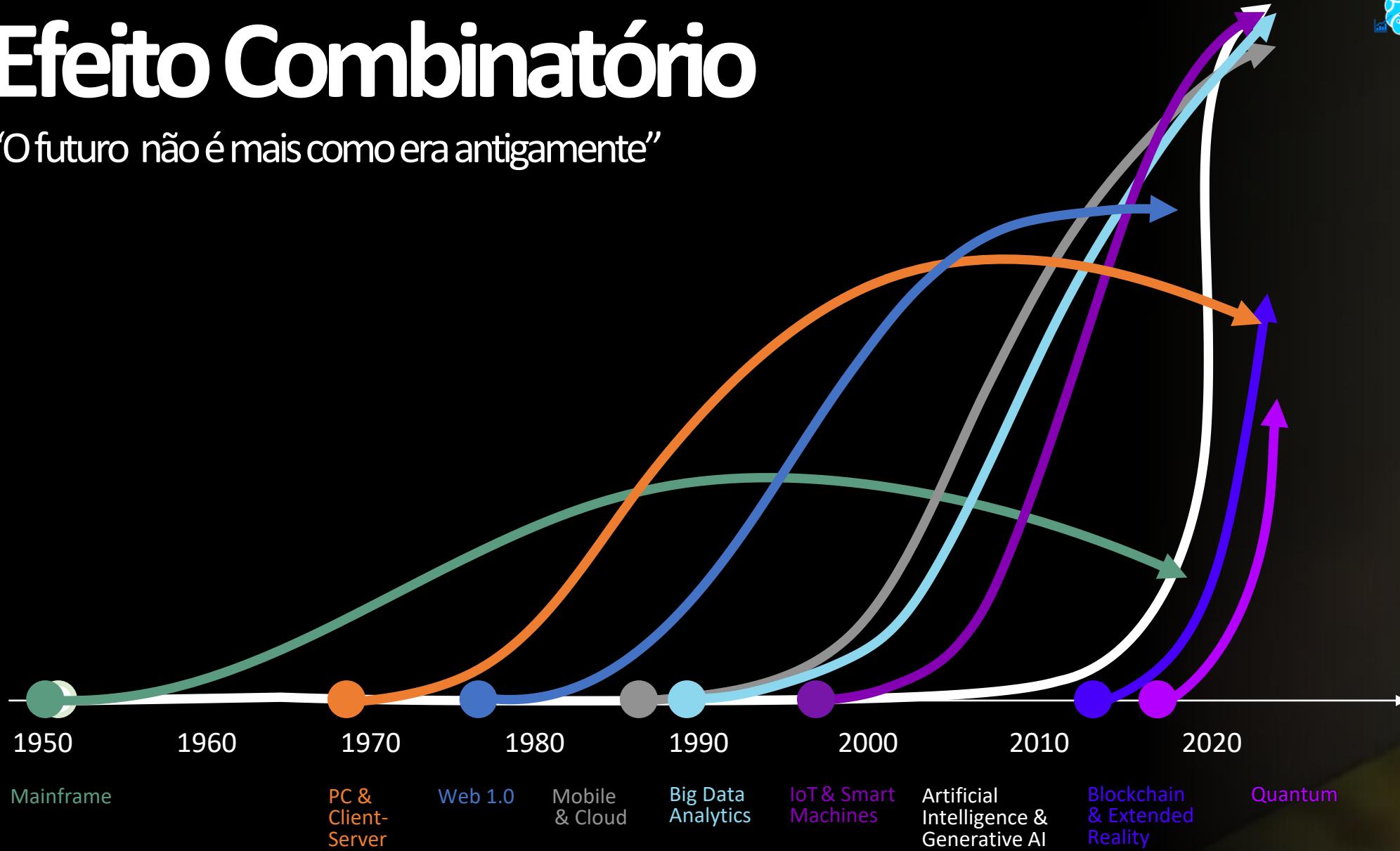
Hype Cycle for Artificial Intelligence, 2024

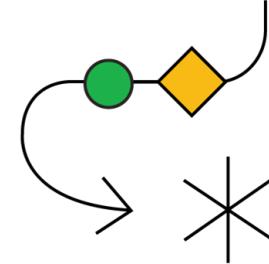


Tecnologias emergindo

# Efeito Combinatório

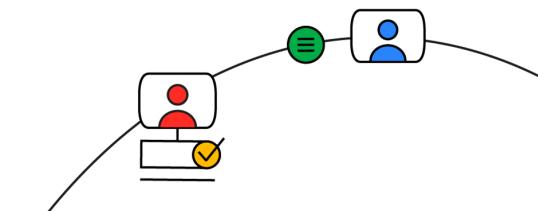
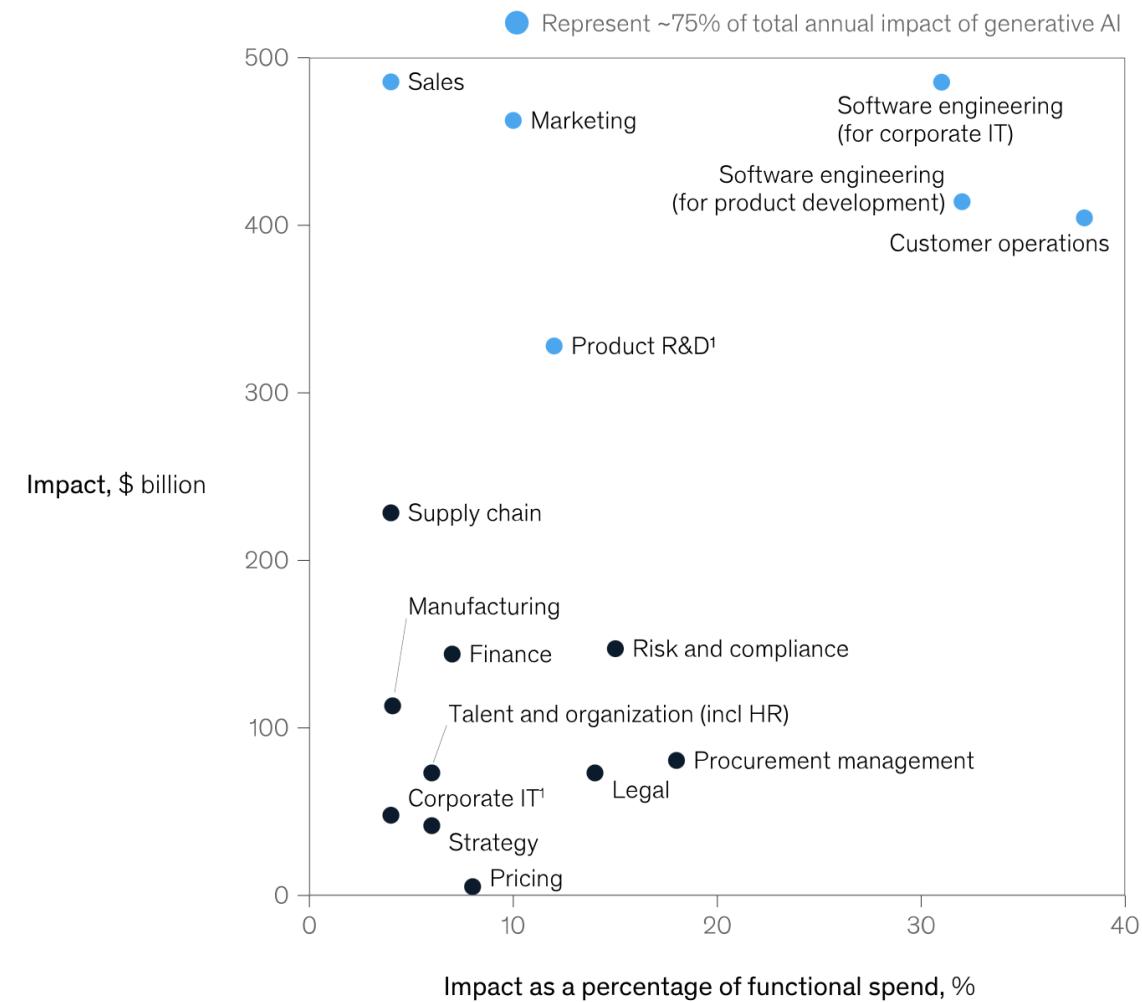
“O futuro não é mais como era antigamente”





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

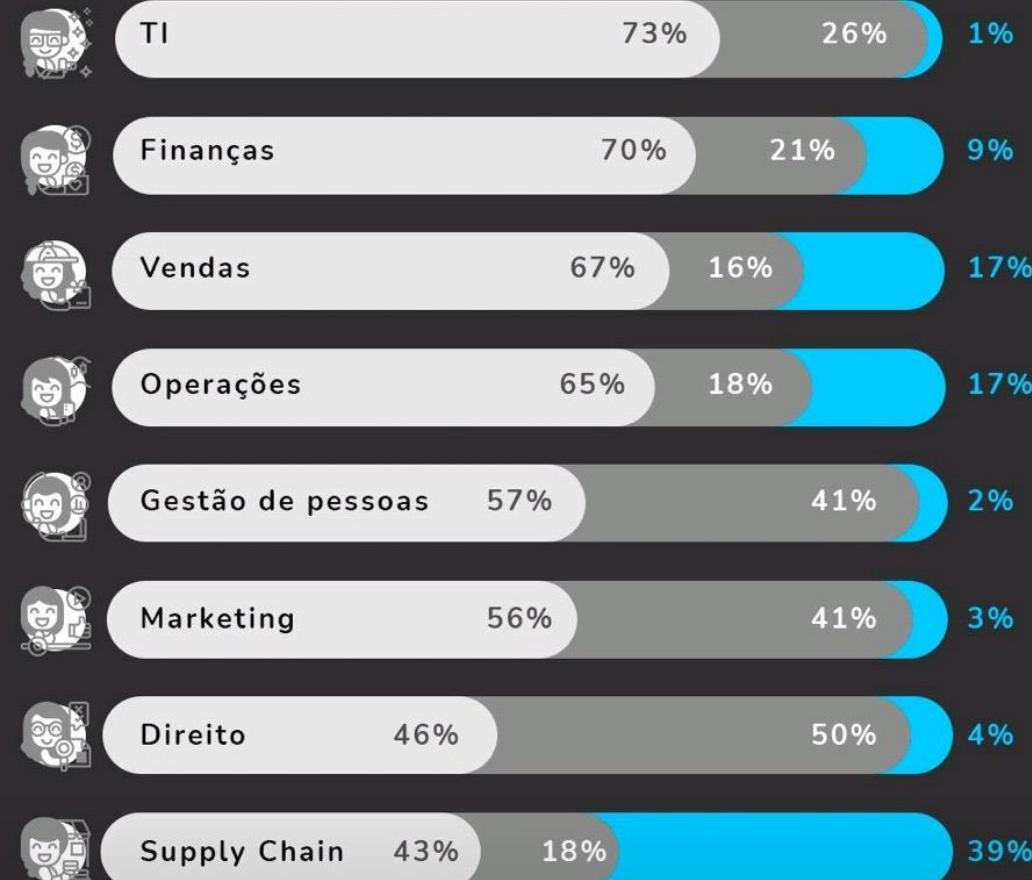
<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/The-economic-potential-of-generative-AI-The-next-productivity-frontier>



Using generative AI  
just a few functions  
could drive major  
technology's  
across potential  
corporate uses

<https://www.mckinsey.com/capabilities/mckinsey-digital/economic-potential-of-generative-AI-The-new-frontier>

## Quem será mais afetado pela IA?

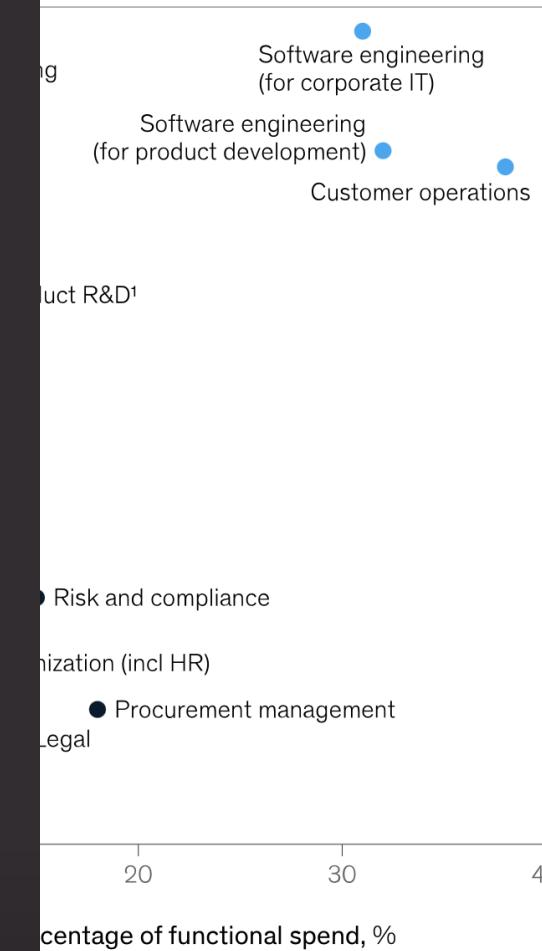


● Grande impacto: tarefas serão automatizadas ou mudarão completamente

● Pequeno impacto: Tarefas não mudam de forma significativa

● Sem impacto

represent ~75% of total annual impact of generative AI



# GenAI na nossa rotina



**CBS NEWS** NEWS SHOWS LIVE LOCAL 🔍

U.S. >

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

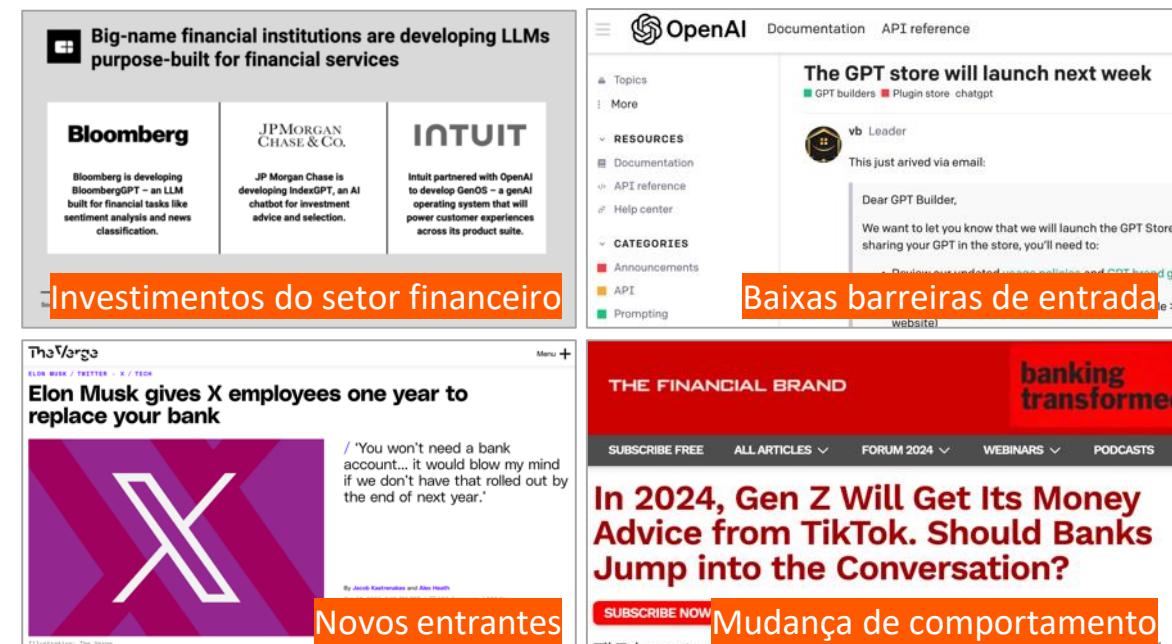
BY ALIZA CHASAN DECEMBER 12, 2023 / 6:45 PM EST / CBS NEWS

The article discusses how dictionary.com chose "hallucinate" as the word of the year, highlighting its rise in usage due to AI.

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**

**The Verge** ELON MUSK / TWITTER / X / TECH

**Elon Musk gives X employees one year to replace your bank**

**Illustration: The Verge**

**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.**

By Jacob Kastrenakes and Alex Heath

**Novos entrantes**

**OpenAI Documentation API reference**

**The GPT store will launch next week**

**GPT builders Plugin store chatgpt**

**vb Leader**

This just arrived 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:

• Review our updated terms of service and GPT builder website

**Baixas barreiras de entrada**

**THE FINANCIAL BRAND**

**banking transformed**

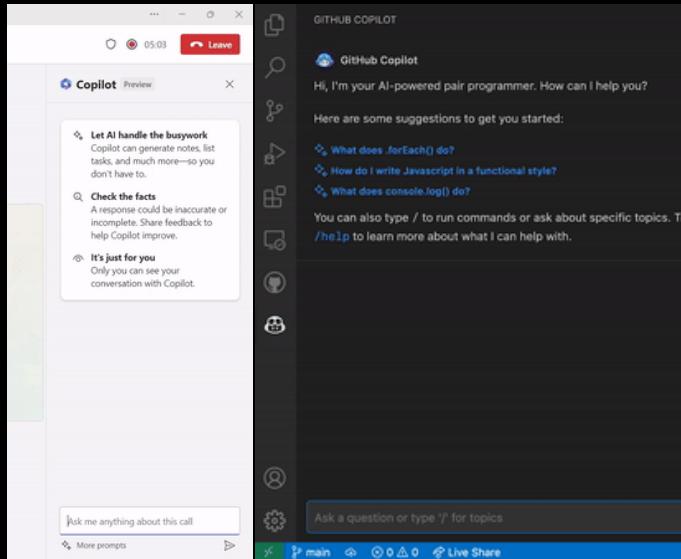
**In 2024, Gen Z Will Get Its Money Advice from TikTok. Should Banks Jump into the Conversation?**

**SUBSCRIBE NOW**

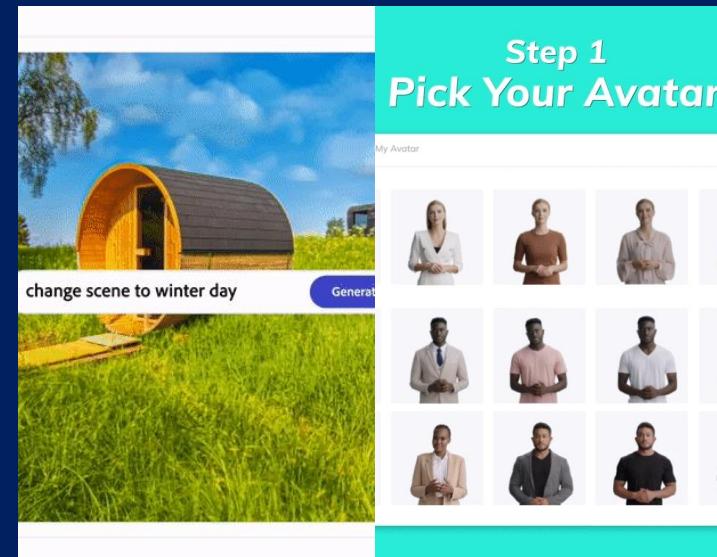
**Mudança de comportamento**

O que é multimodal?

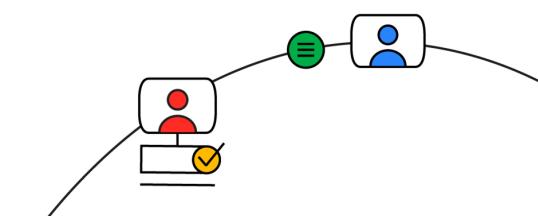
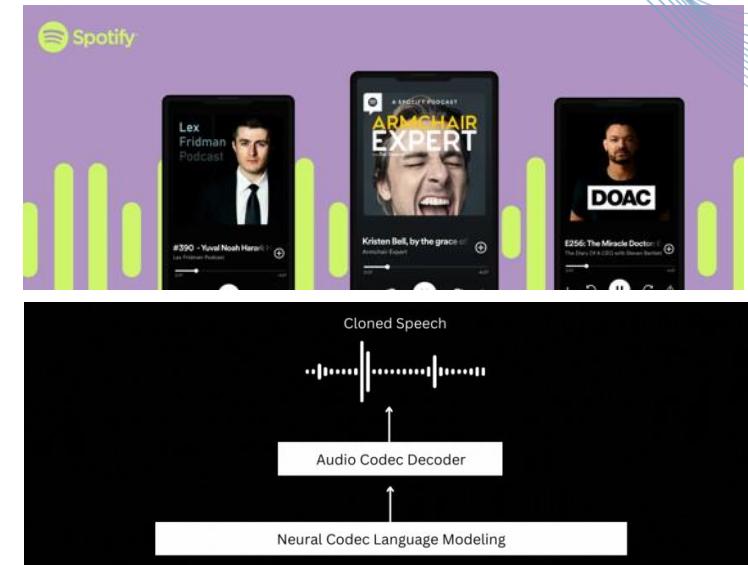
# Textos & Código

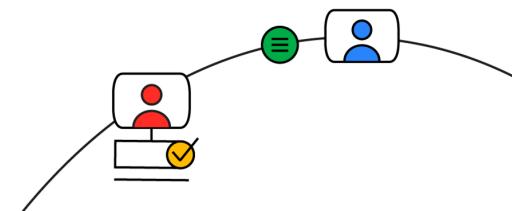


# Imagens, Vídeos



# Áudio

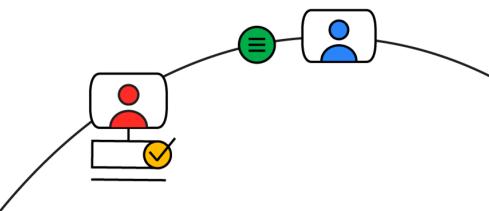


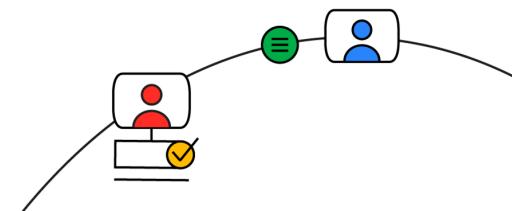
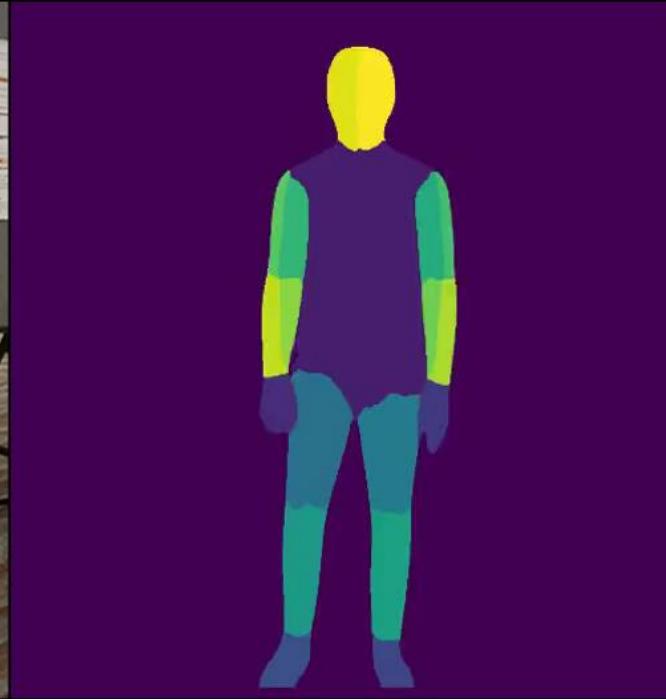
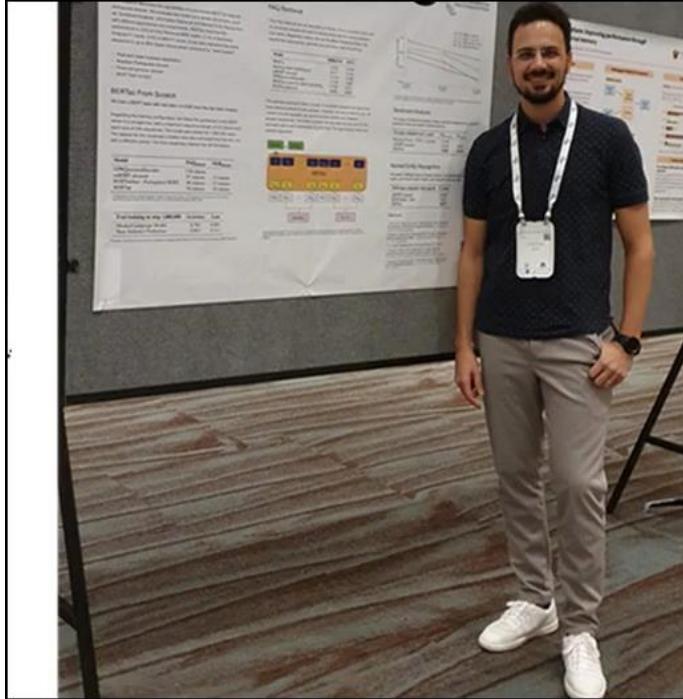


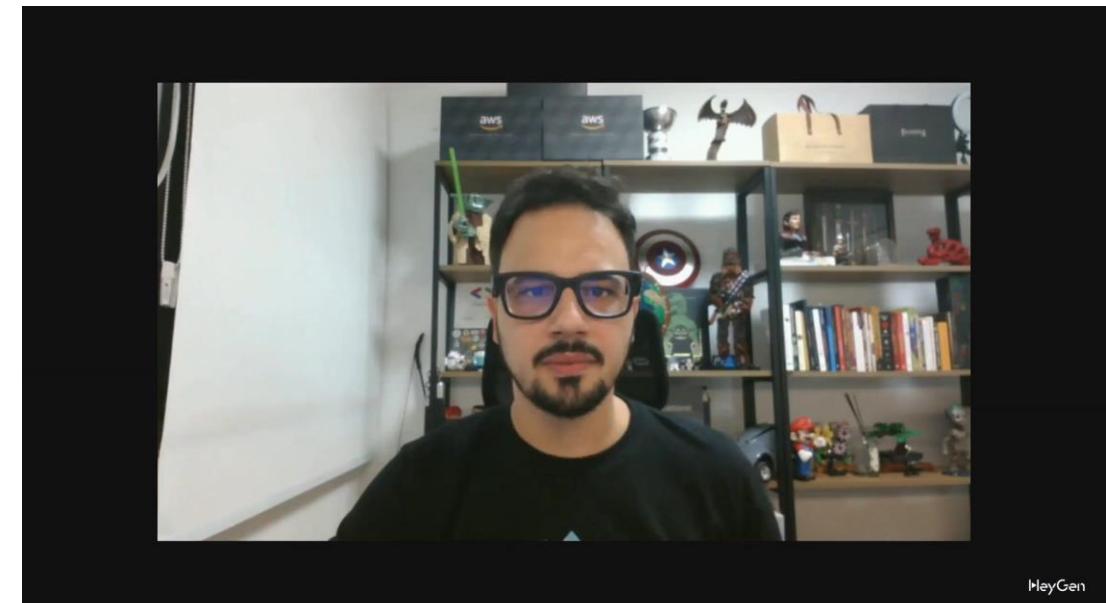
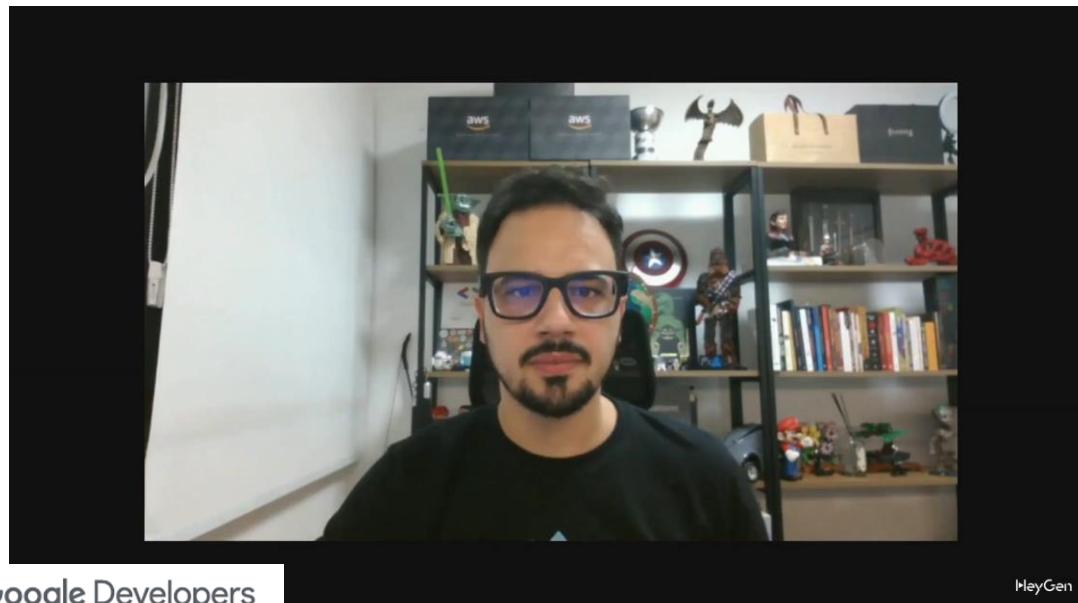


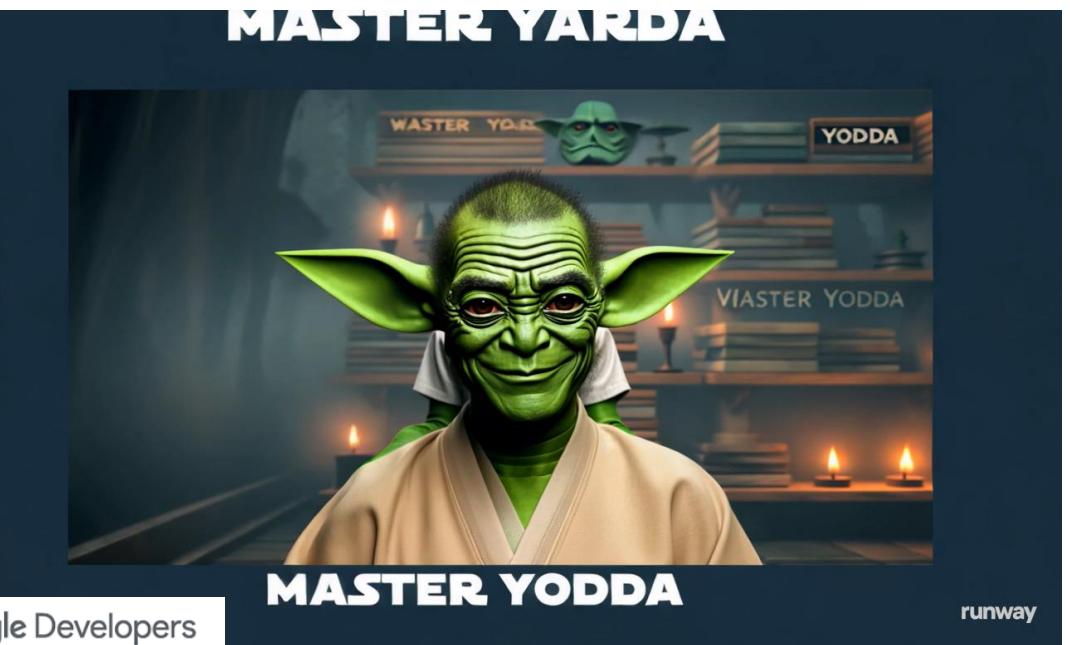


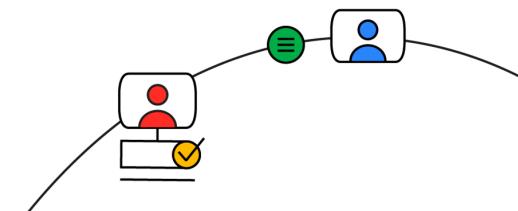
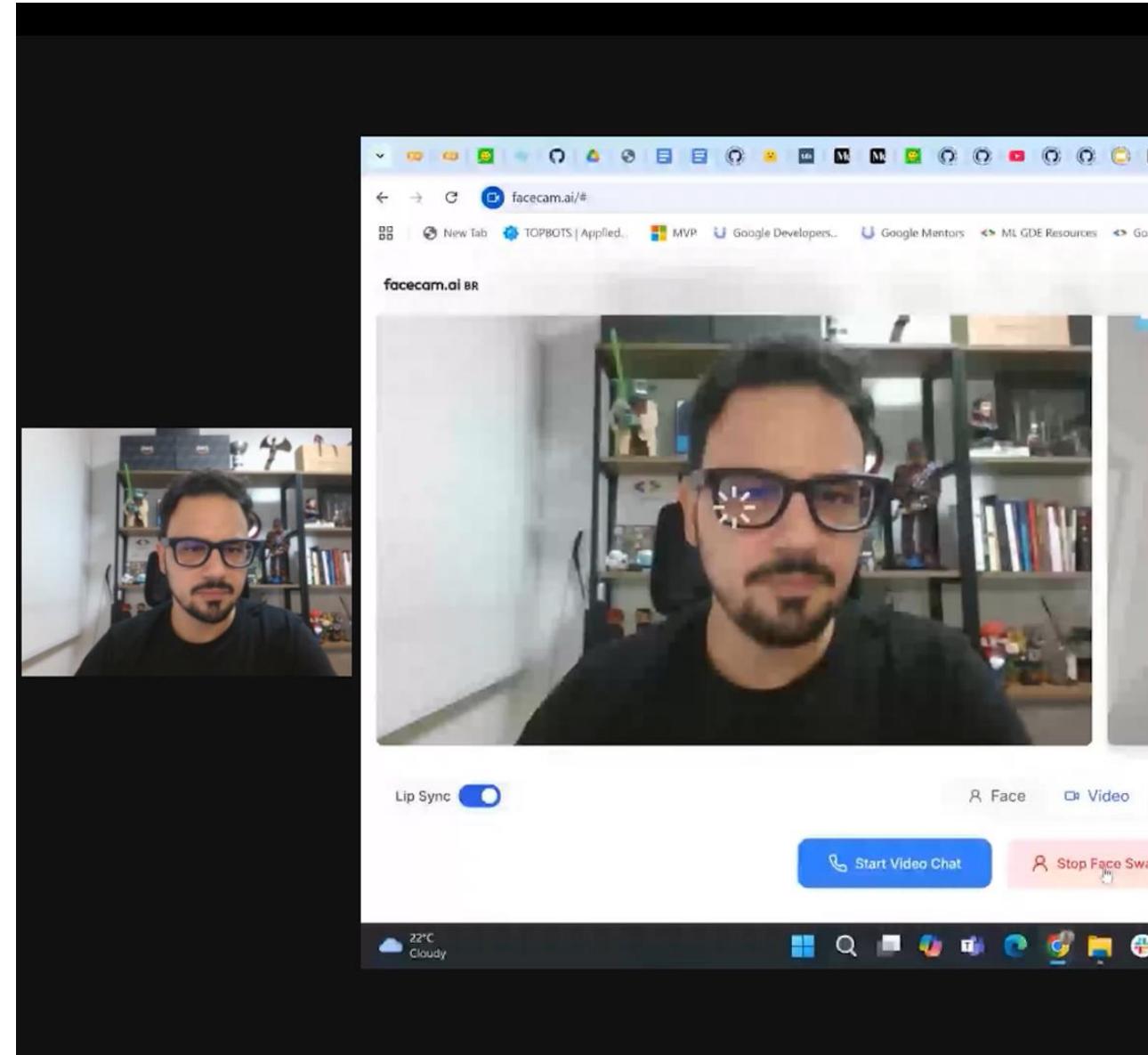
🤯 This AI can replace ANYONE from a video! Movie sets will become a thing of the past...

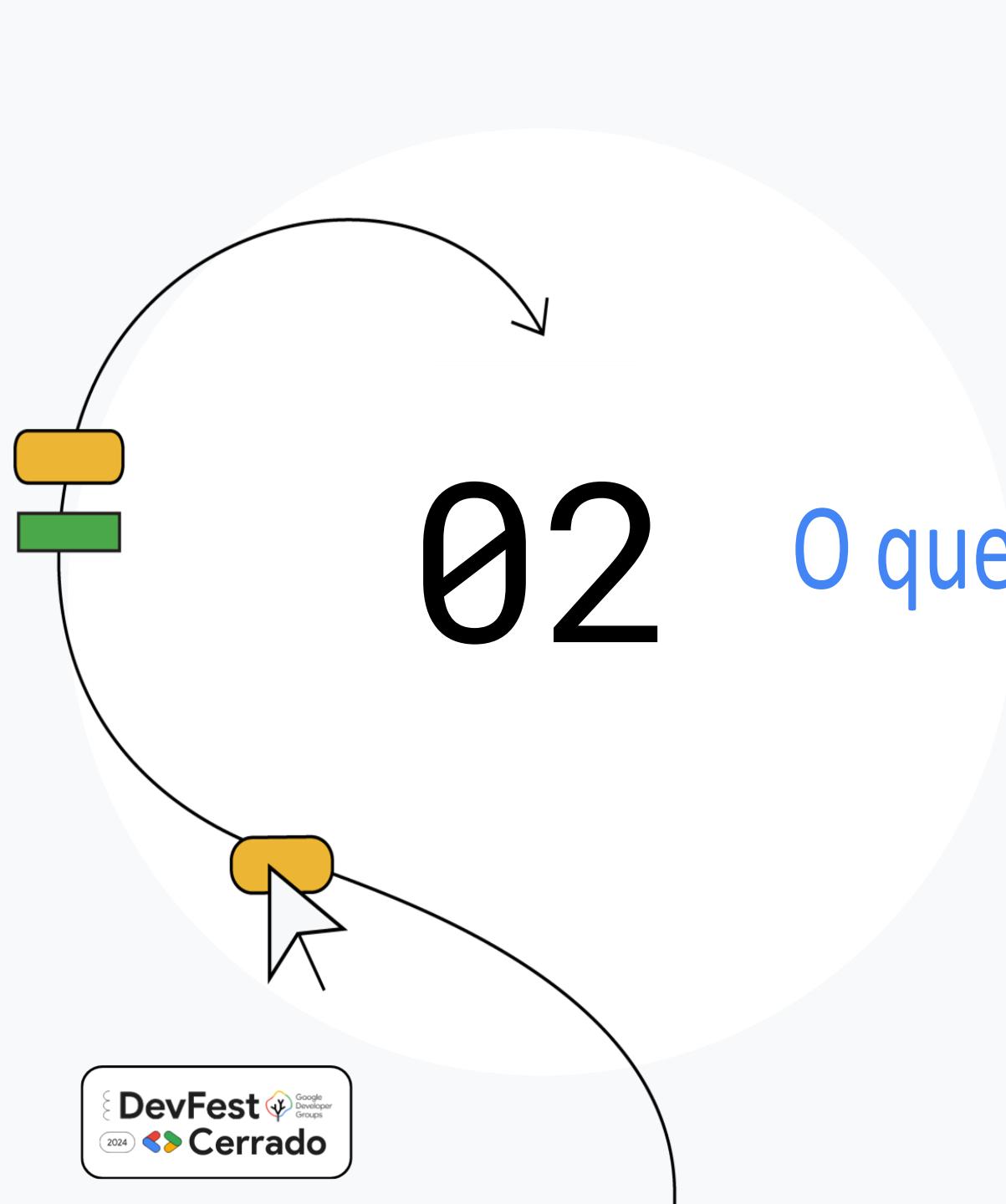






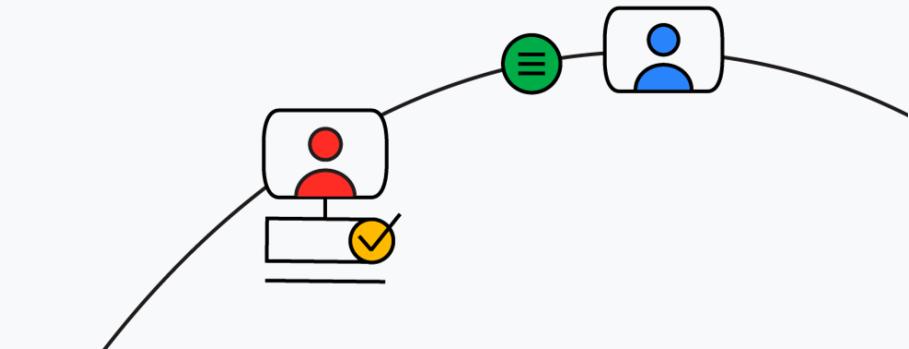






02

# O que tem por trás da GenAI?



# Como representar contexto/significado das palavras

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



# Como representar contexto/significado das palavras

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



# Como representar contexto/significado das palavras

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

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



# Como representar contexto/significado das palavras

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

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

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

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

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 o cérebro faz isso?



# 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.



# 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

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

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

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

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

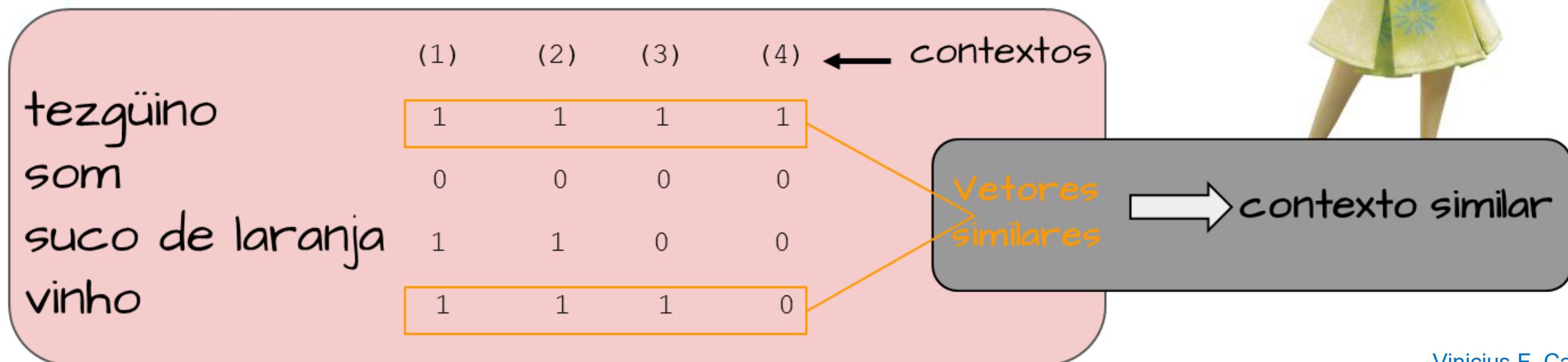
Vetores similares



# Como representar contexto/significado das palavras

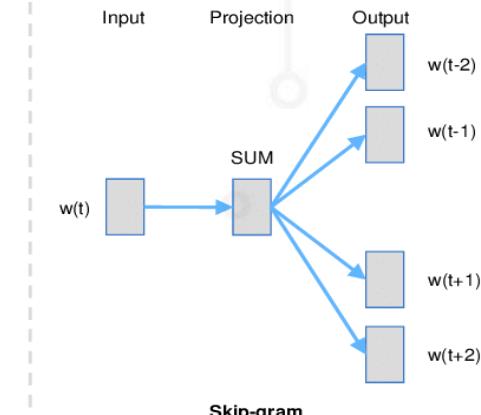
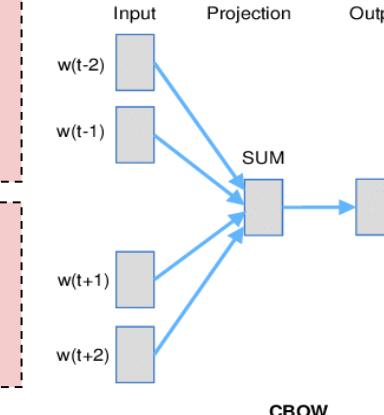
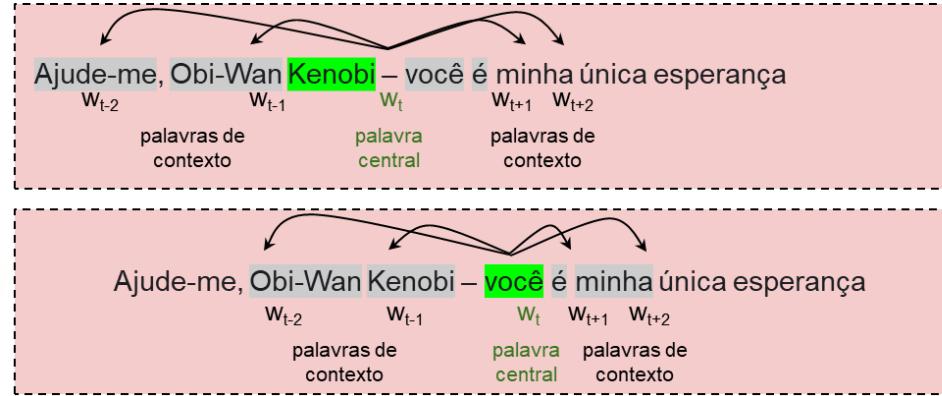
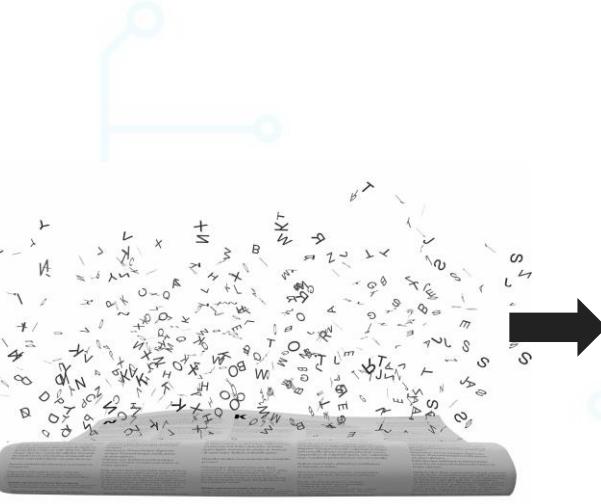
Inserindo contexto de forma manual...

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.



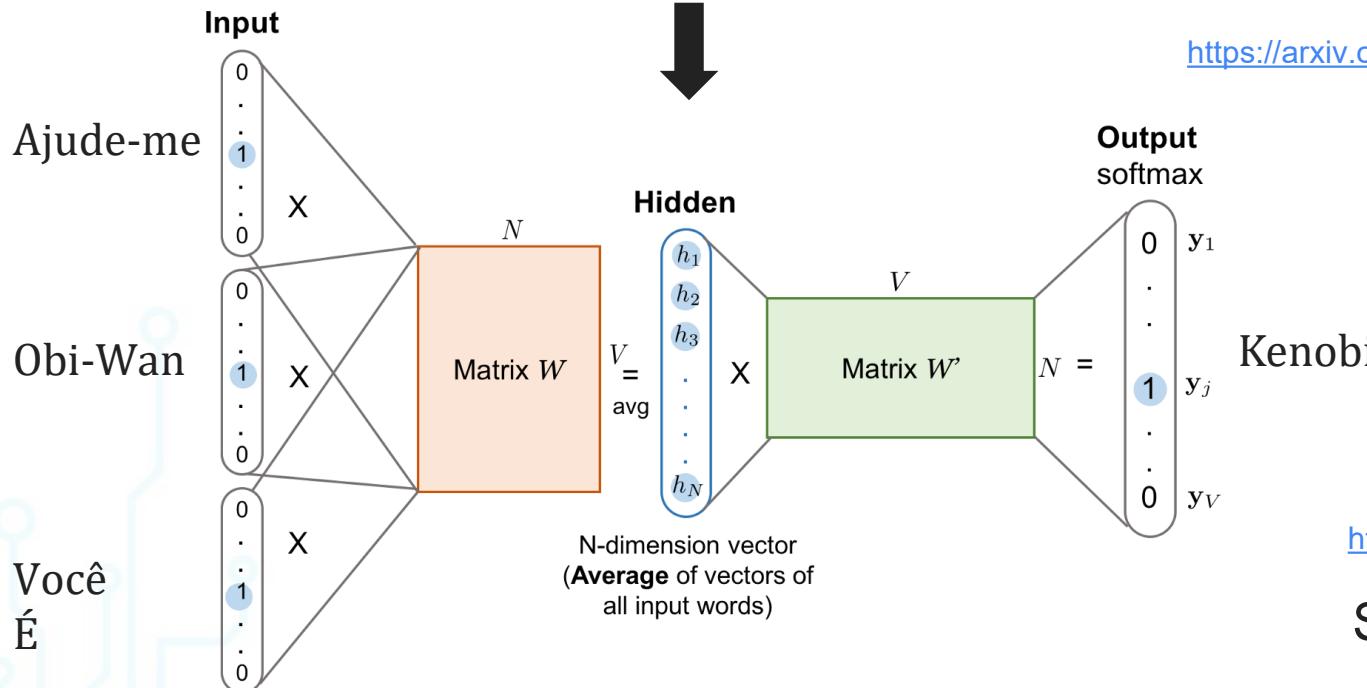
# Processamento de Linguagem Natural

## Natural Language Processing (NLP)



<https://arxiv.org/abs/1301.3781>

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



MultiLayer Perceptron MLP

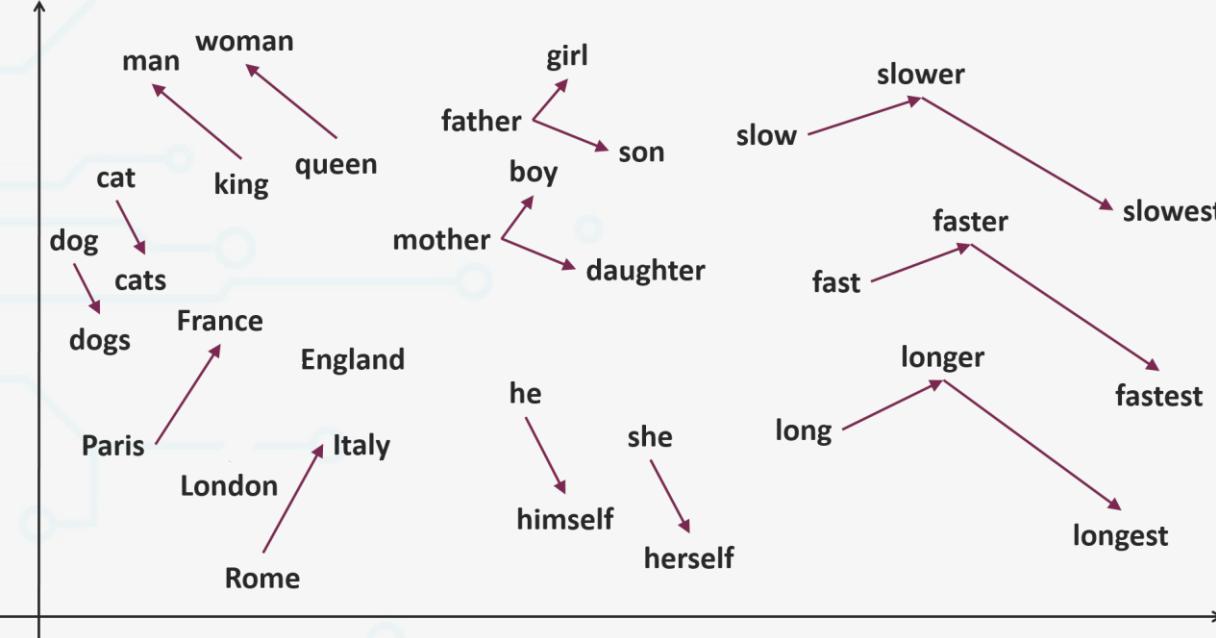
<https://dl.acm.org/doi/10.5555/1639537.1639542>

Self Supervised Learning (SSL)

<https://arxiv.org/abs/2110.09327>

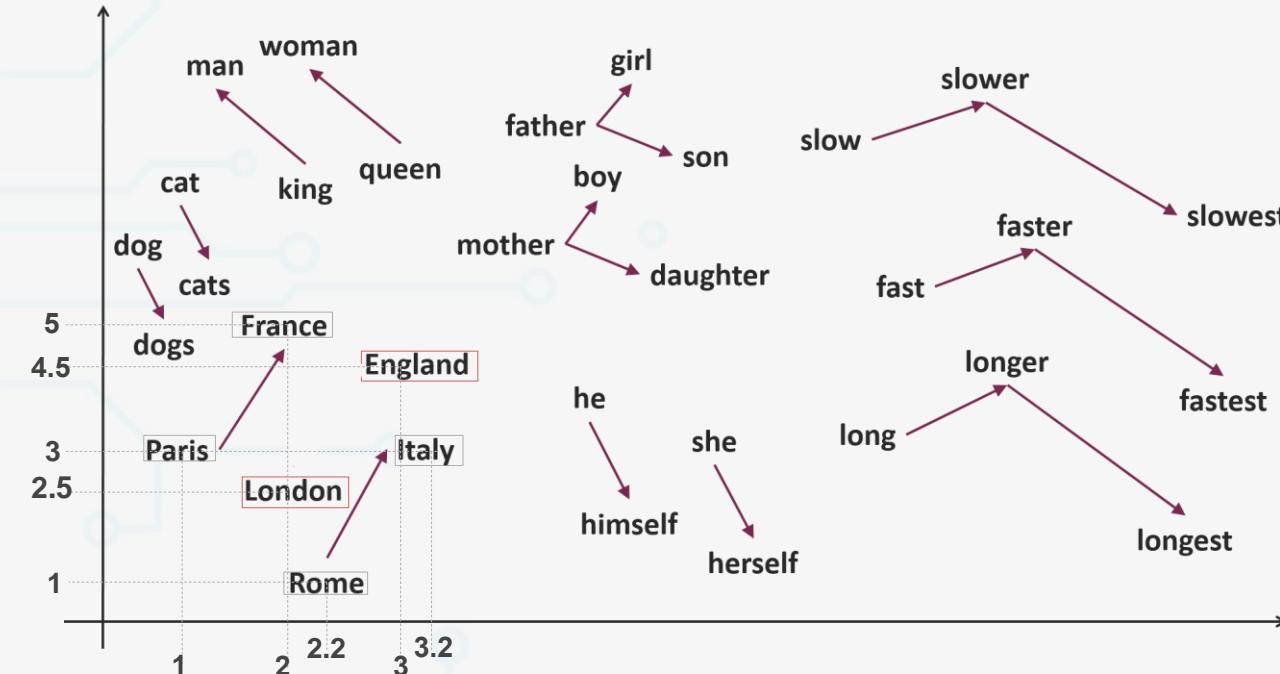
# Processamento de Linguagem Natural

## Natural Language Processing (NLP)



# Processamento de Linguagem Natural

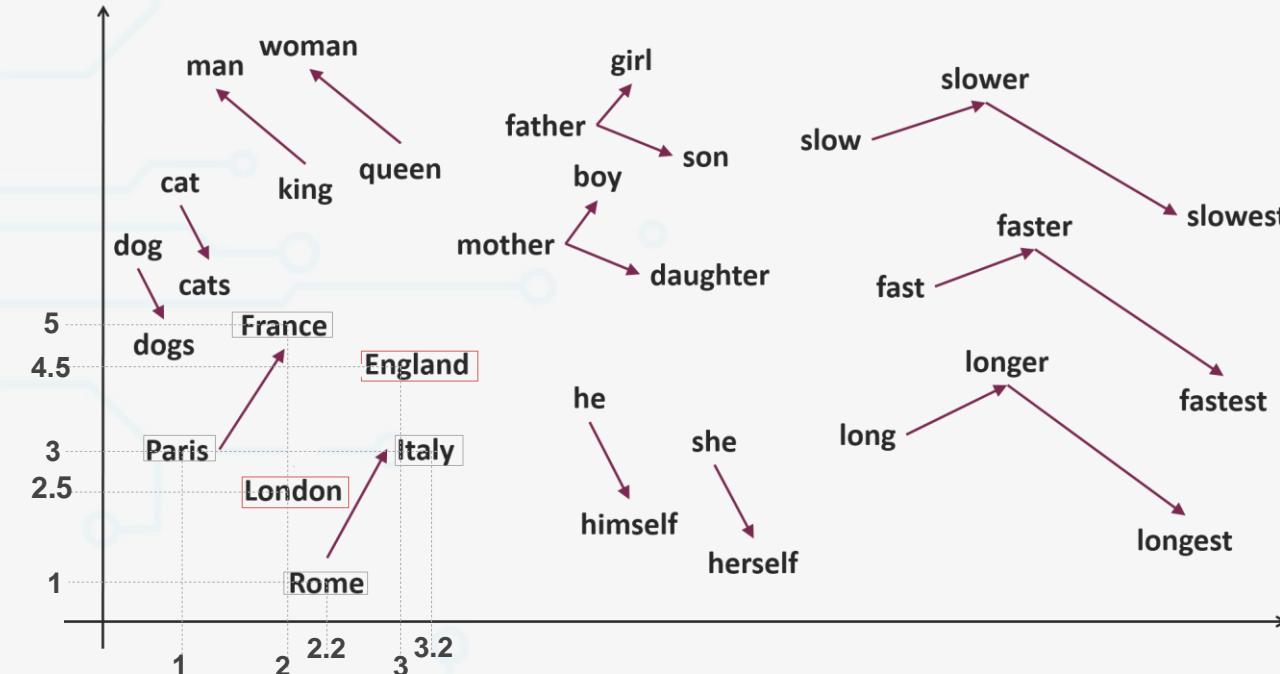
## Natural Language Processing (NLP)



- 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

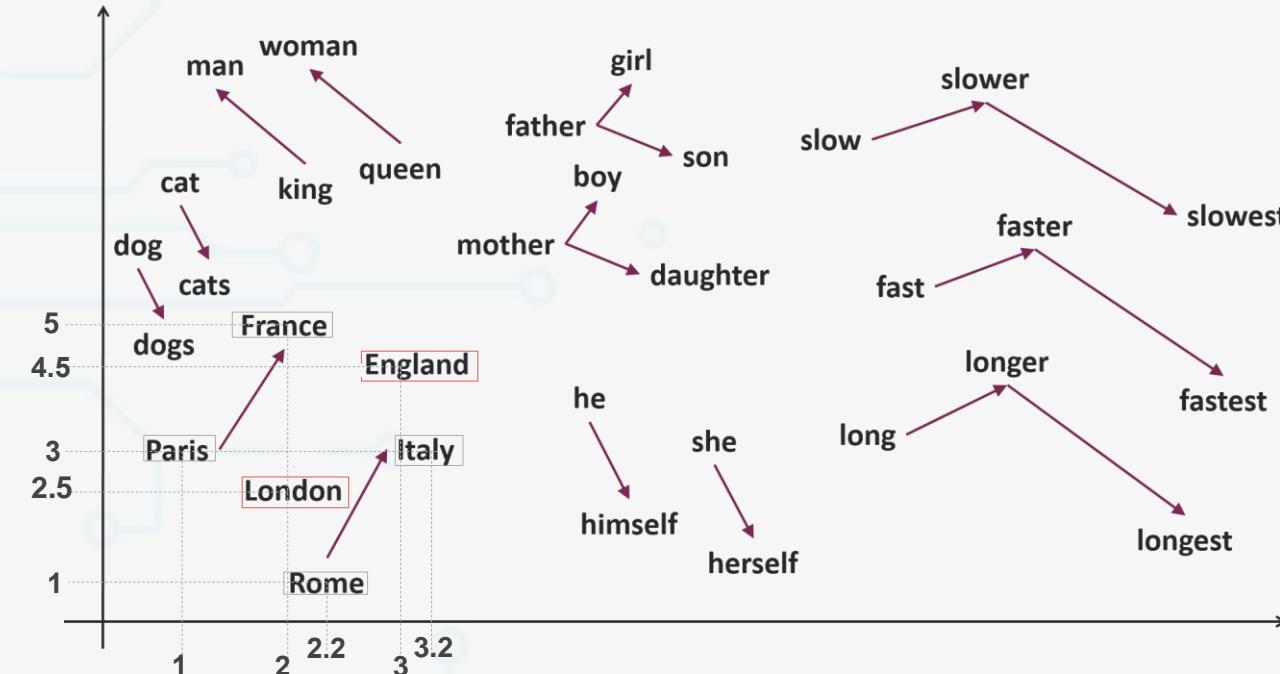
## Natural Language Processing (NLP)



Qual a capital da Inglaterra?

# Processamento de Linguagem Natural

## Natural Language Processing (NLP)

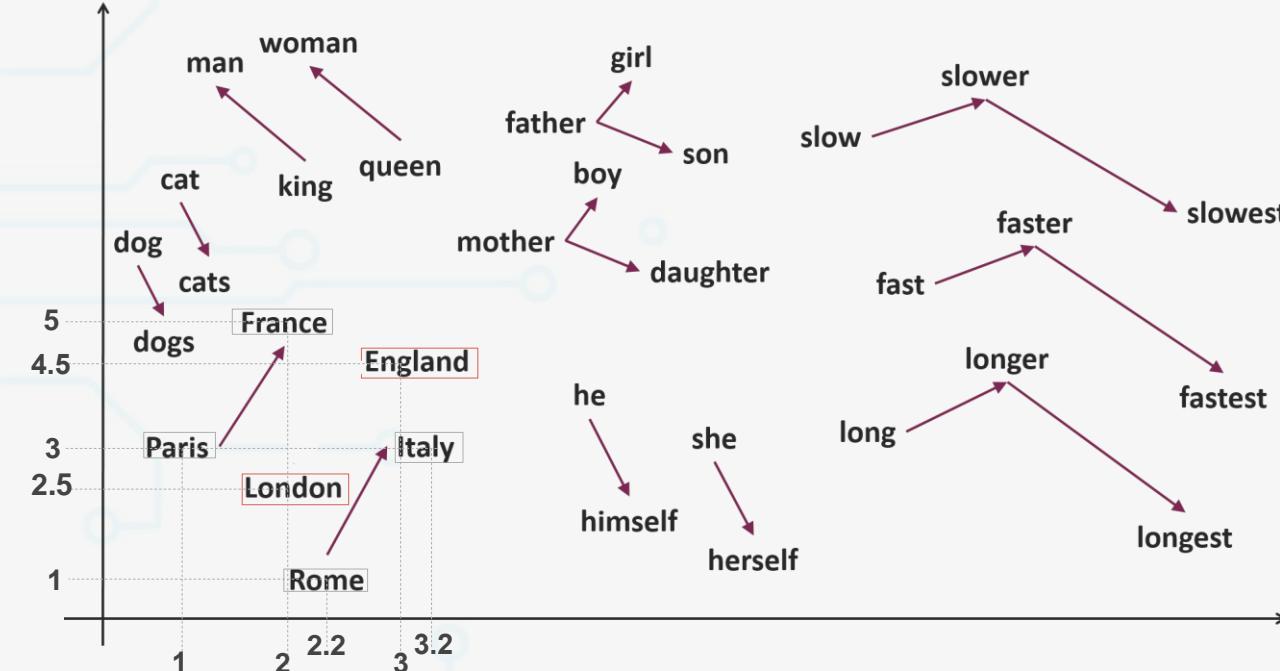


Qual a capital da Inglaterra?

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]

=

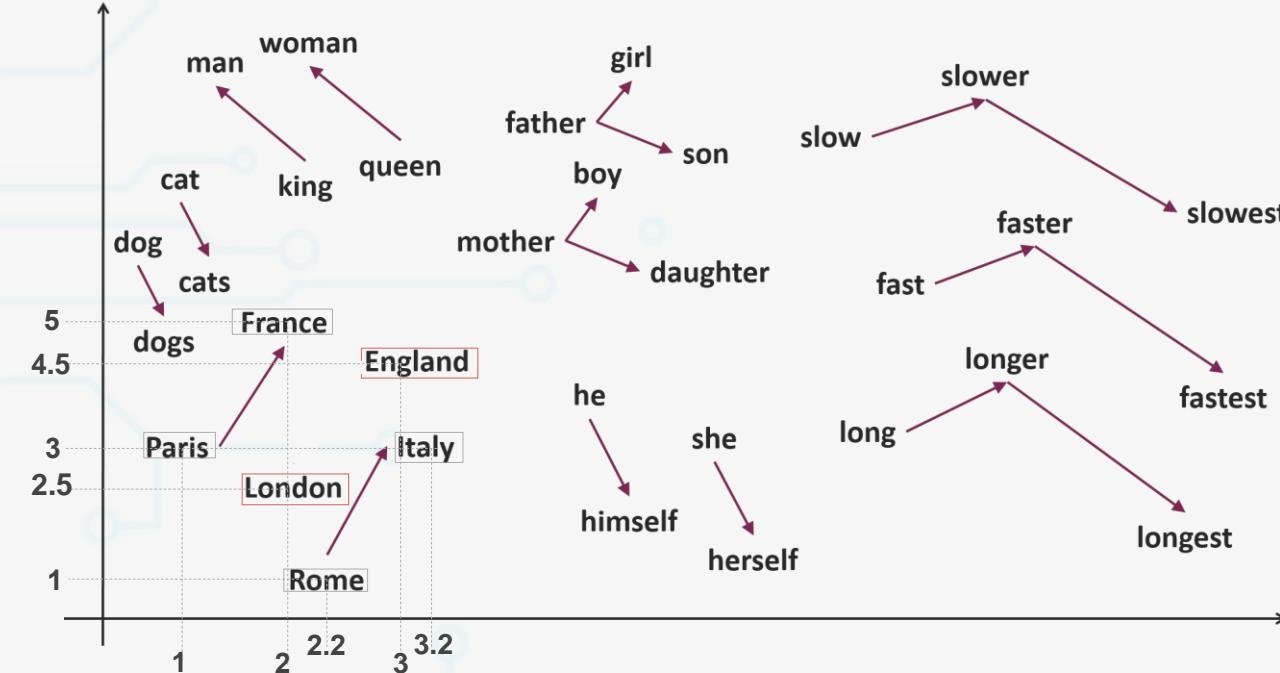
Result. [2, 2.5]

==

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]

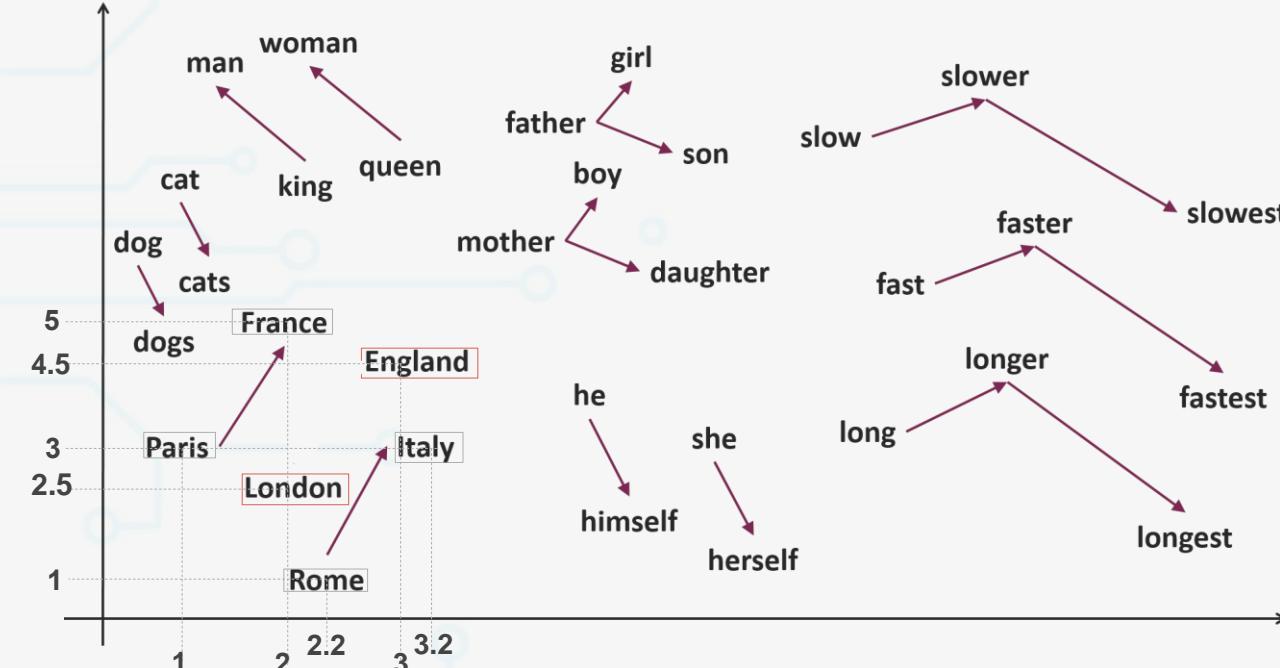
==

London [2, 2.5]

Paris – France + England = London

# Processamento de Linguagem Natural

## Natural Language Processing (NLP)



Qual a capital da Inglaterra?

$$\begin{aligned}
 \text{Paris} & [1, 3] \\
 \text{France} & [2, 5] \\
 = & \\
 \text{Result.} & [-1, -2] \\
 + & \\
 \text{England} & [3, 4.5] \\
 = & \\
 \text{Result.} & [2, 2.5] \\
 == & \\
 \text{London} & [2, 2.5]
 \end{aligned}$$

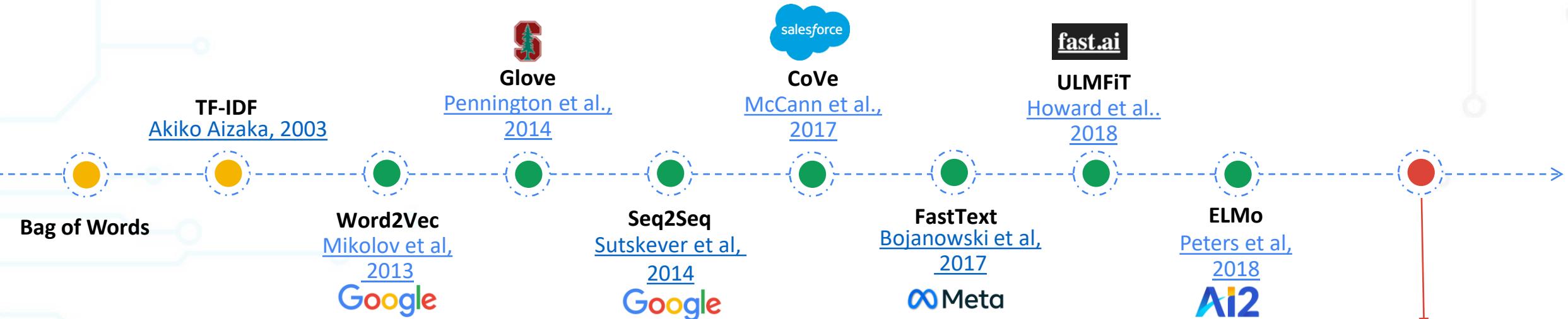
Paris – France + England = London

$$\begin{aligned}
 \text{Rome} & [2.2, 1] \\
 \text{Italy} & [3.2, 3] \\
 = & \\
 \text{Result.} & [-1, -2] \\
 + & \\
 \text{England} & [3, 4.5] \\
 = & \\
 \text{Result.} & [2, 2.5] \\
 == & \\
 \text{London} & [2, 2.5]
 \end{aligned}$$

Rome – Italy + England = London

# Processamento de Linguagem Natural

## Natural Language Processing (NLP)



### 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

Lukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\* ‡  
illia.polosukhin@gmail.com

Google

<https://arxiv.org/abs/1706.03762>

## 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

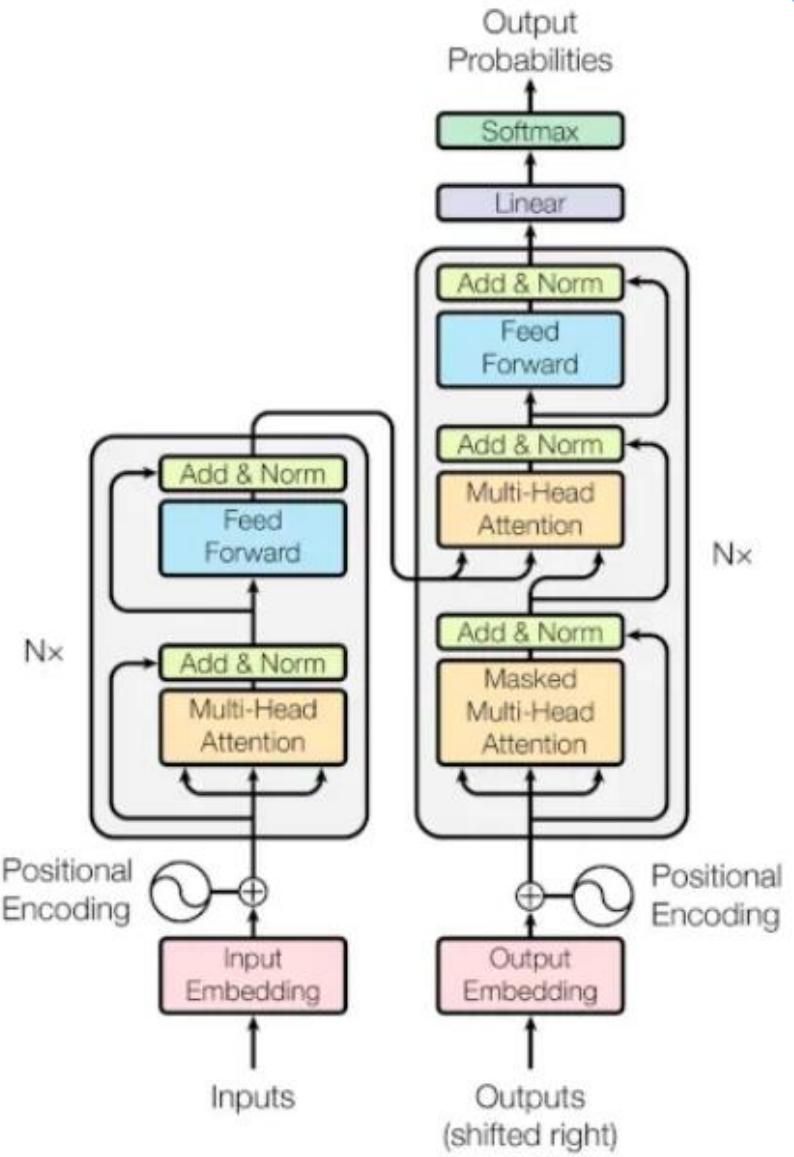
**Lukasz Kaiser\***  
Google Brain  
lukasz.kaiser@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

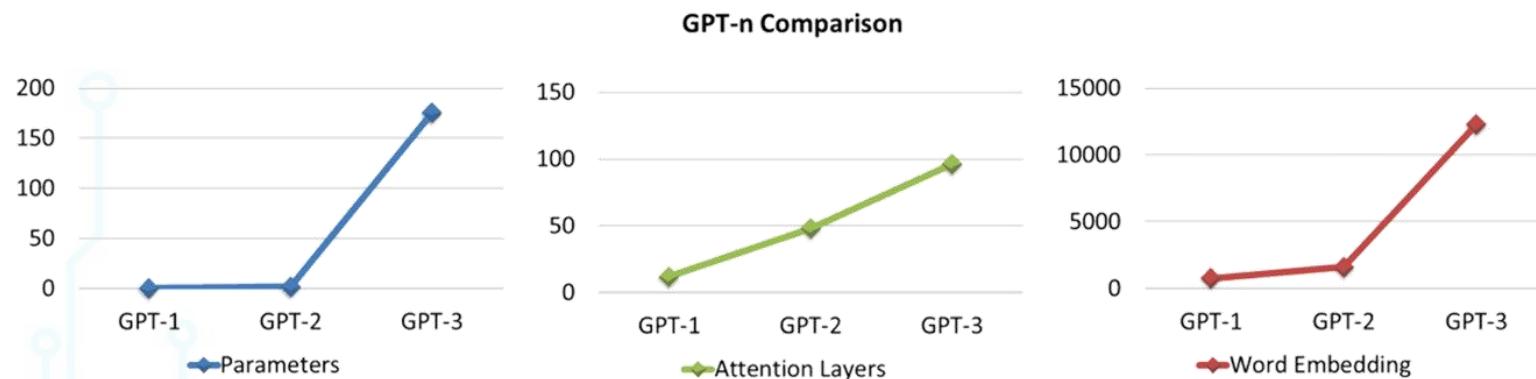
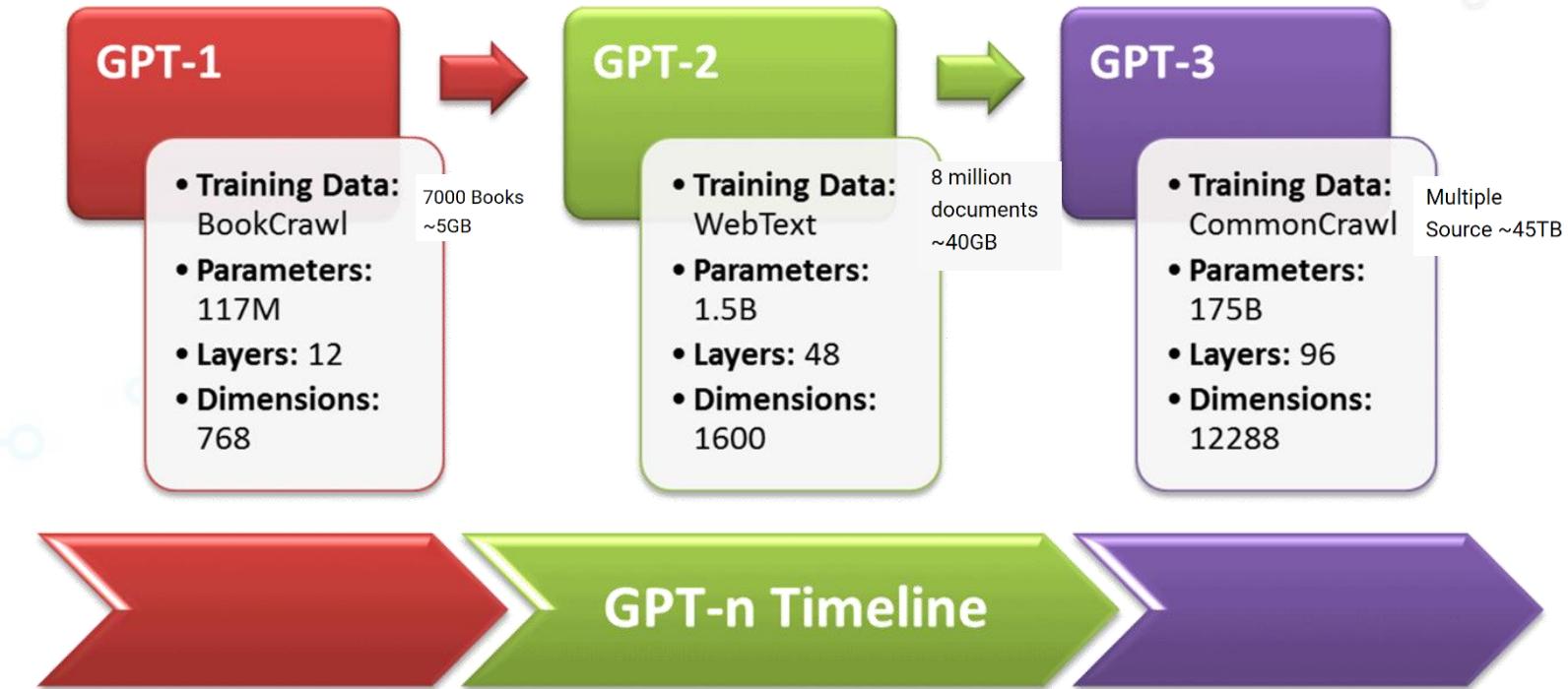
### Abstract

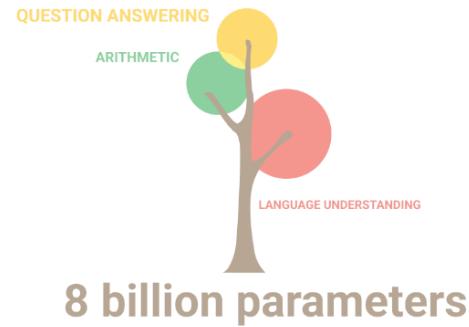
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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, our model establishes a new single model state-of-the-art BLEU score of 41.9 after

<https://arxiv.org/abs/1706.03762>



# Família GPT



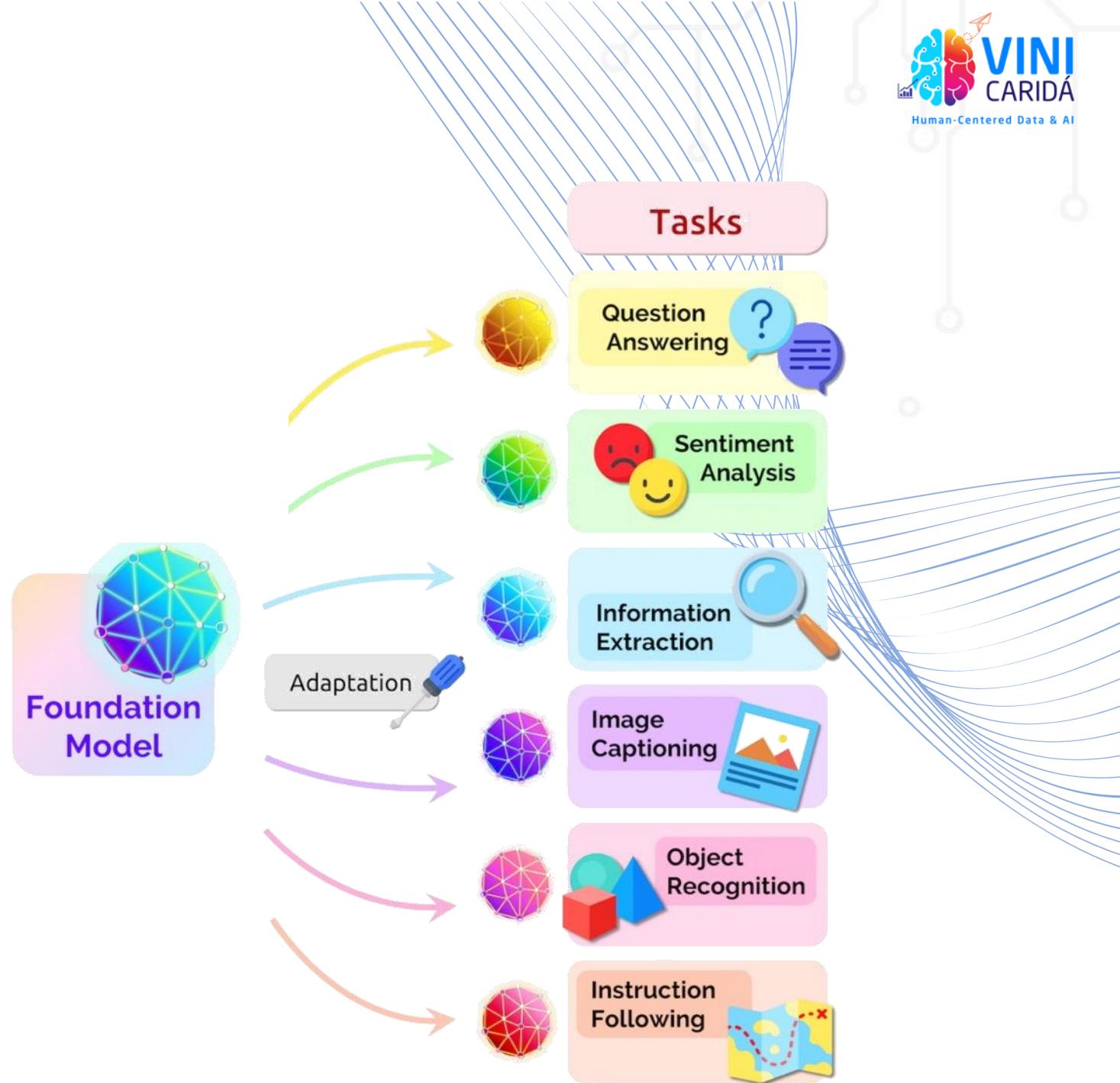


## Zero-Shot Learning in Modern NLP

<https://joeddav.github.io/blog/2020/05/29/ZSL.html>

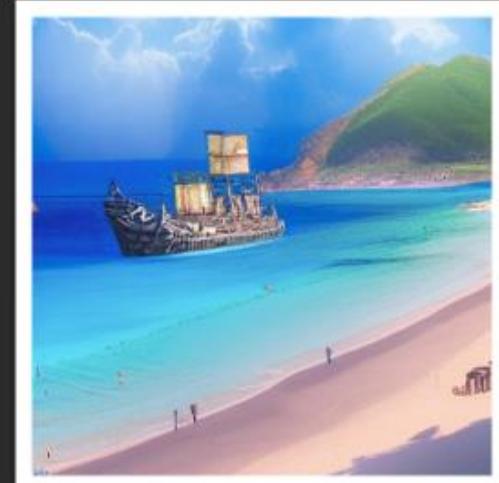
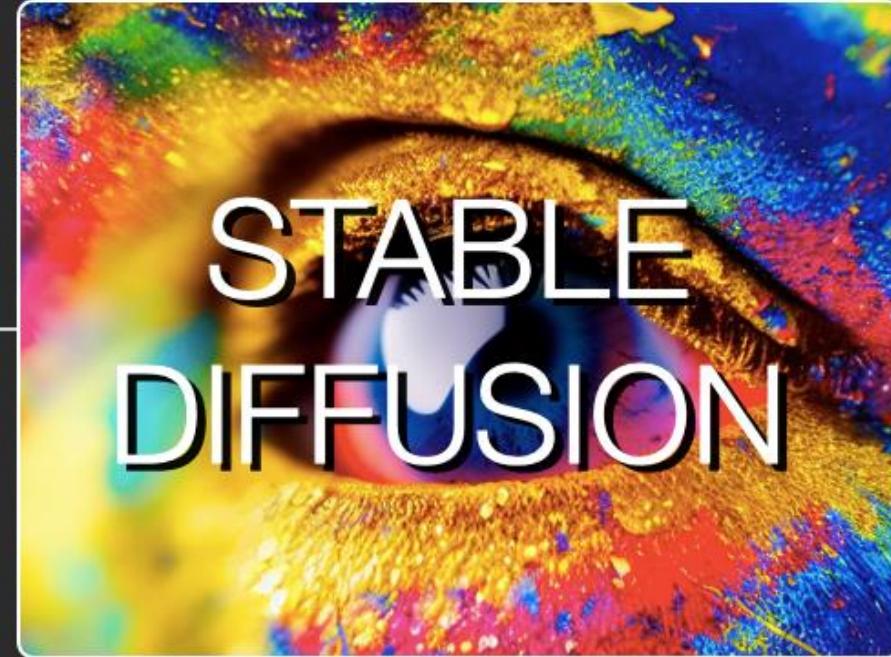
## Few-Shot Learning

Não Corporativo <https://arxiv.org/pdf/1904.05046.pdf>



Attention Is All You Need

Pirate ship



<https://arxiv.org/abs/1706.03762>

Experts

Google Developers

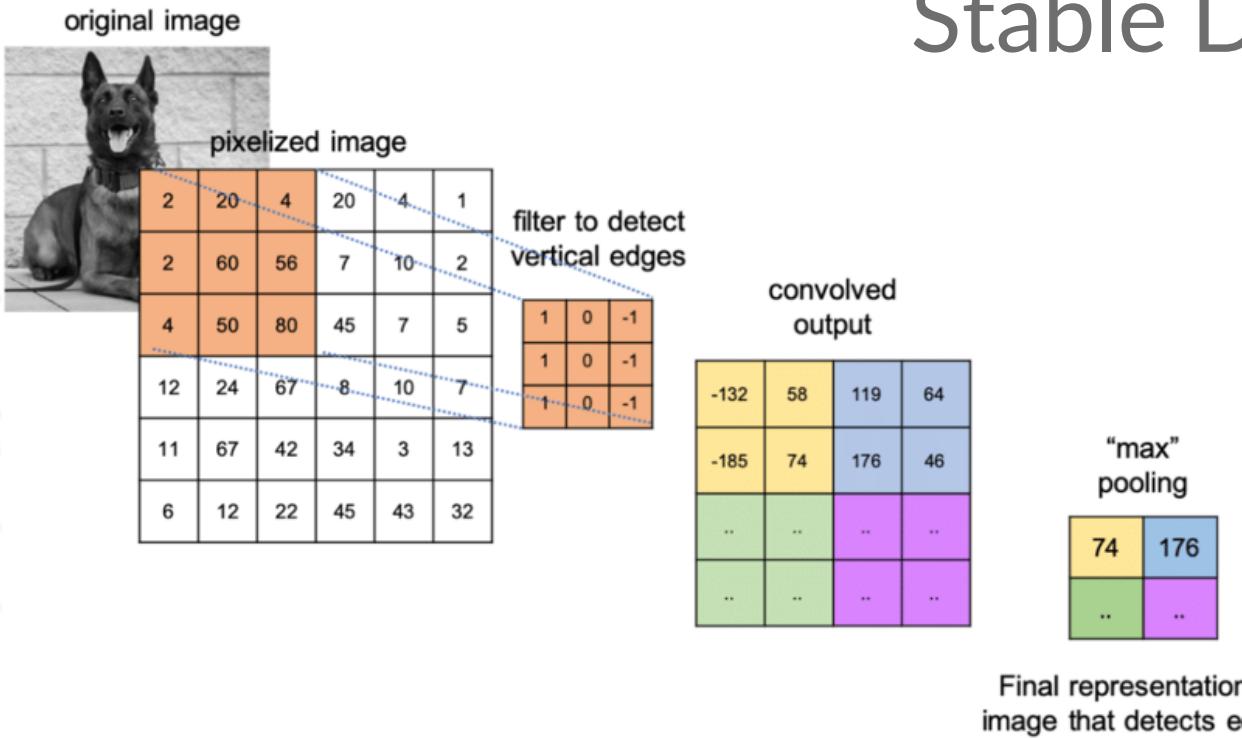
Inputs

Outputs  
(shifted right)

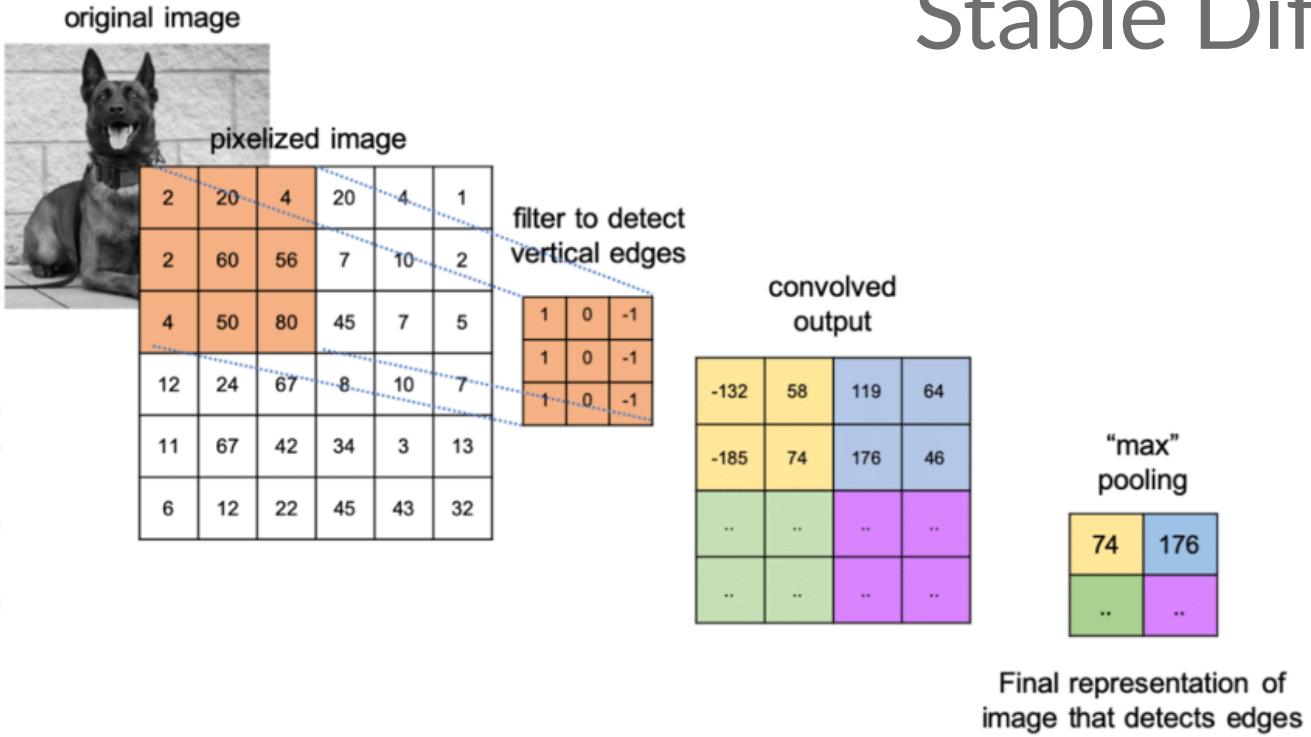
Embedding

Embedding

# Stable Diffusion

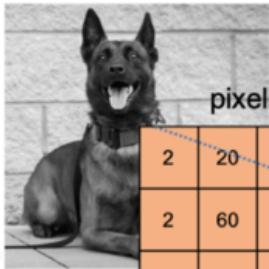


# Stable Diffusion



# Stable Diffusion

original image



pixelized image

2	20	4	20	4	1
2	60	56	7	10	2
4	50	80	45	7	5
12	24	67	8	10	7
11	67	42	34	3	13
6	12	22	45	43	32

filter to detect  
vertical edges

1	0	-1
1	0	-1
1	0	-1

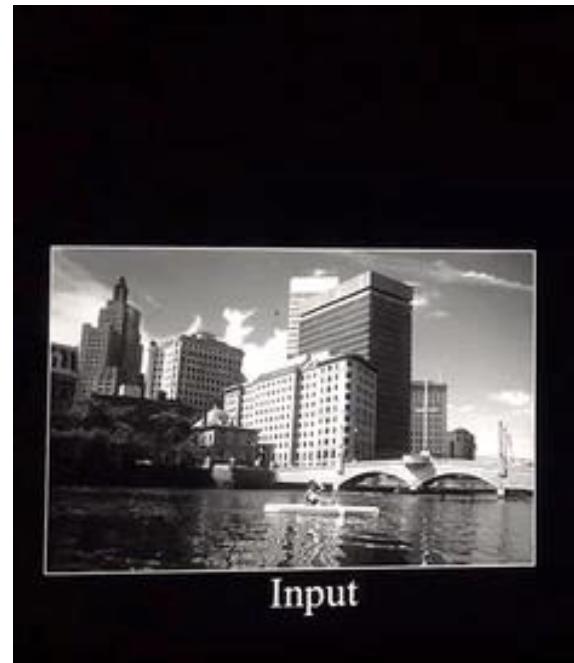
convolved  
output

-132	58	119	64
-185	74	176	46
..	..	..	..
74	176	..	..

"max"  
pooling

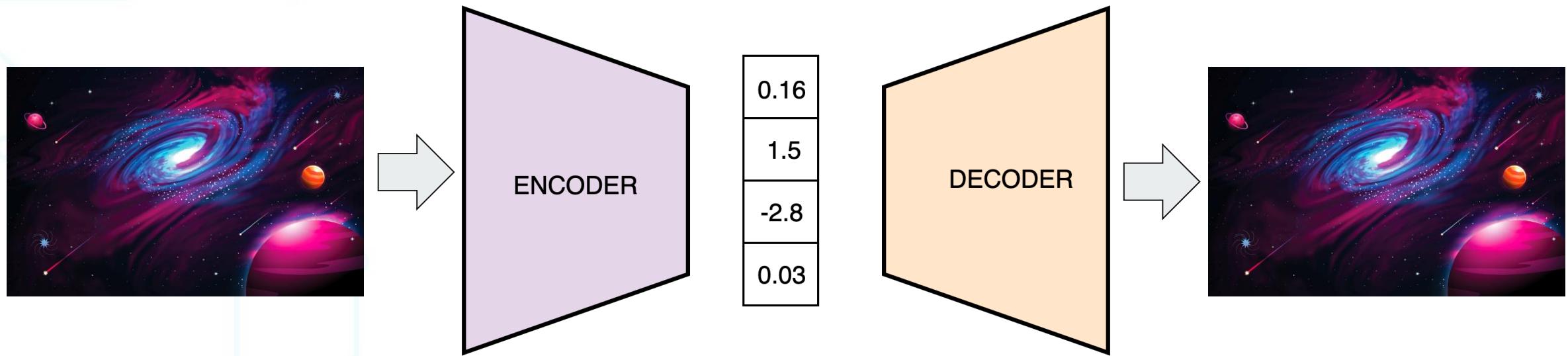
74	176
..	..

Final representation of  
image that detects edges



# Stable Diffusion

## Espaço latente



# Modelos de difusão

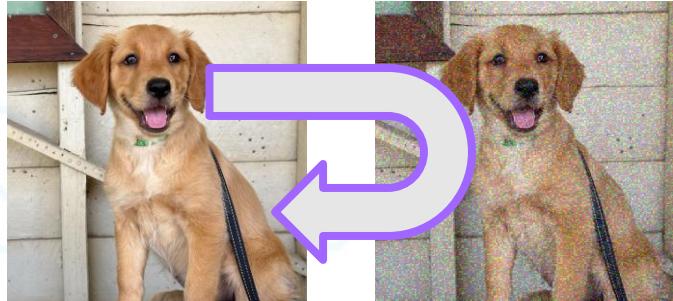
## Diffusion model



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

# Modelos de difusão

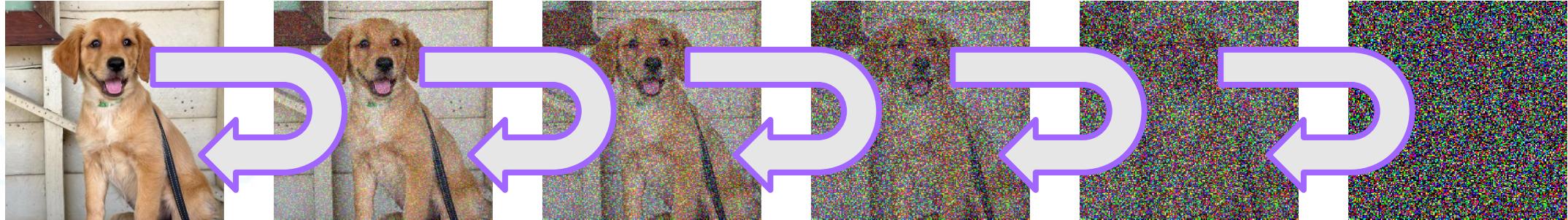
## Diffusion model



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

# Modelos de difusão

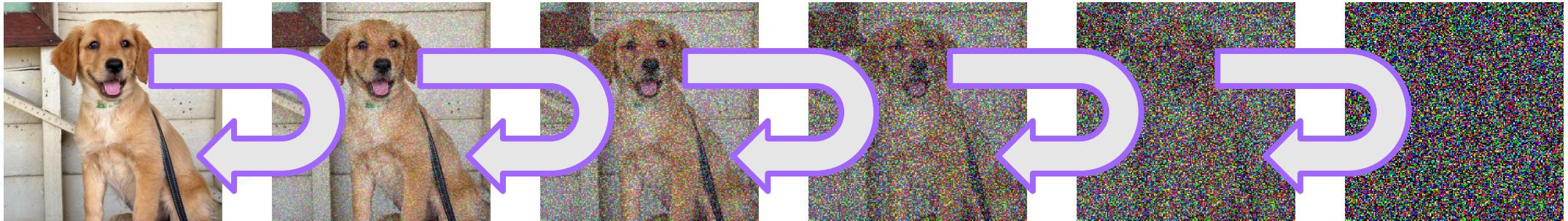
## Diffusion model



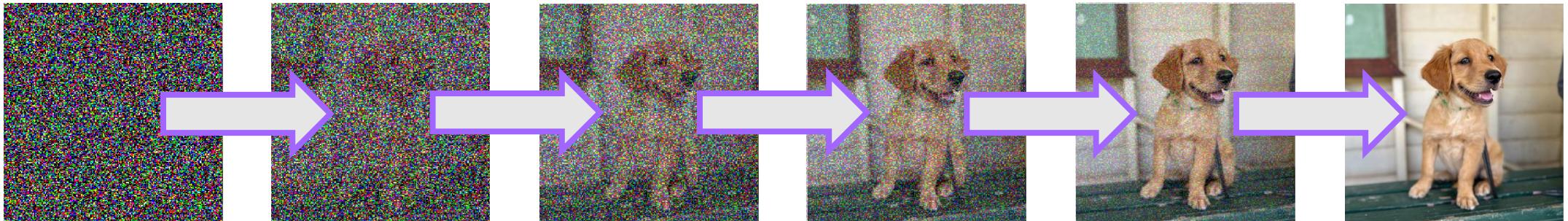
Adds noise and learns how to work backwards to the original image.

# Modelos de difusão

## Diffusion model

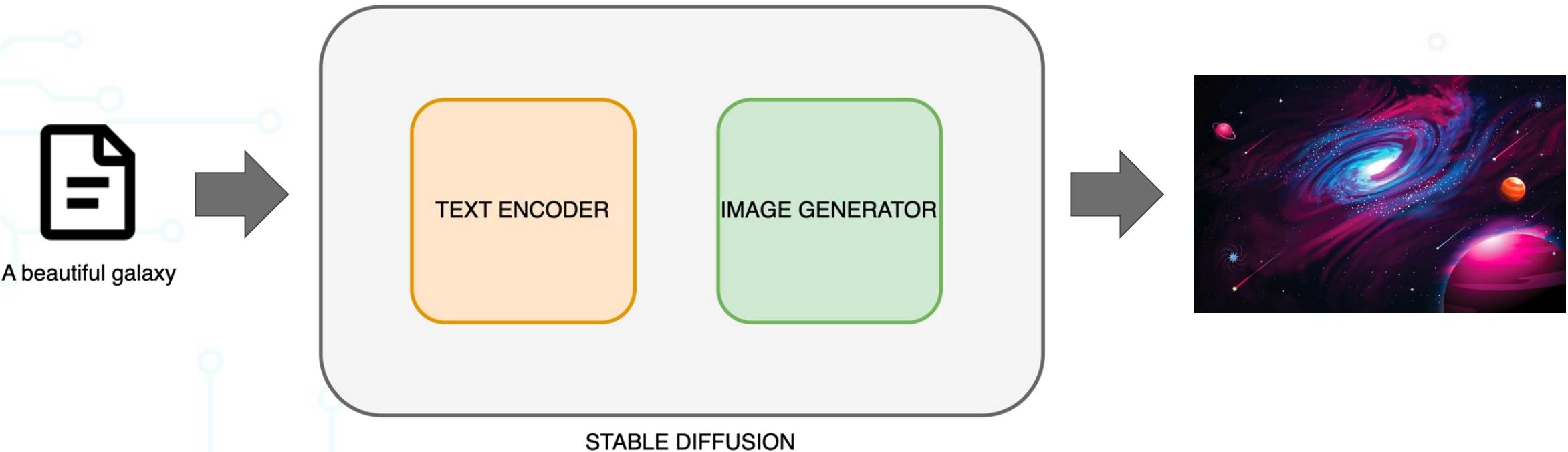


Adds noise and learns how to work backwards to the original image.



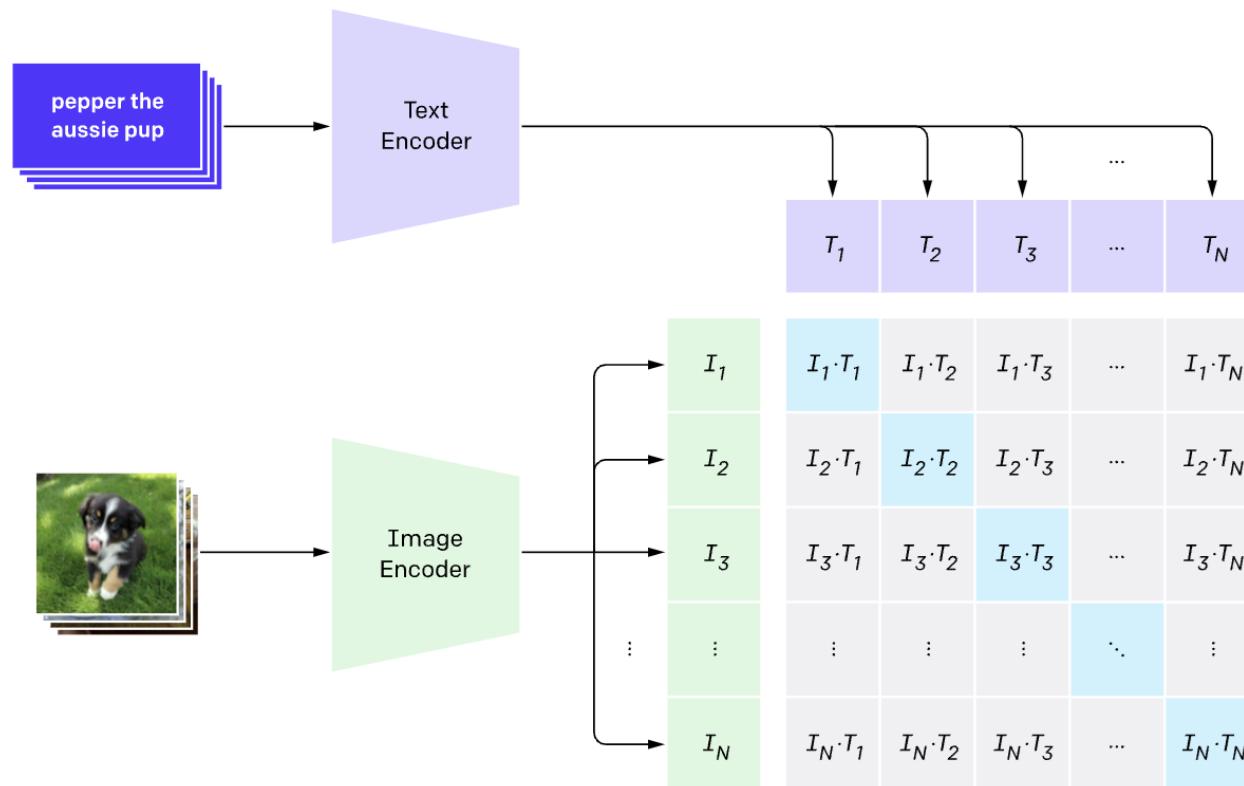
Trained model works from random noise to generated an image.

# Stable Diffusion



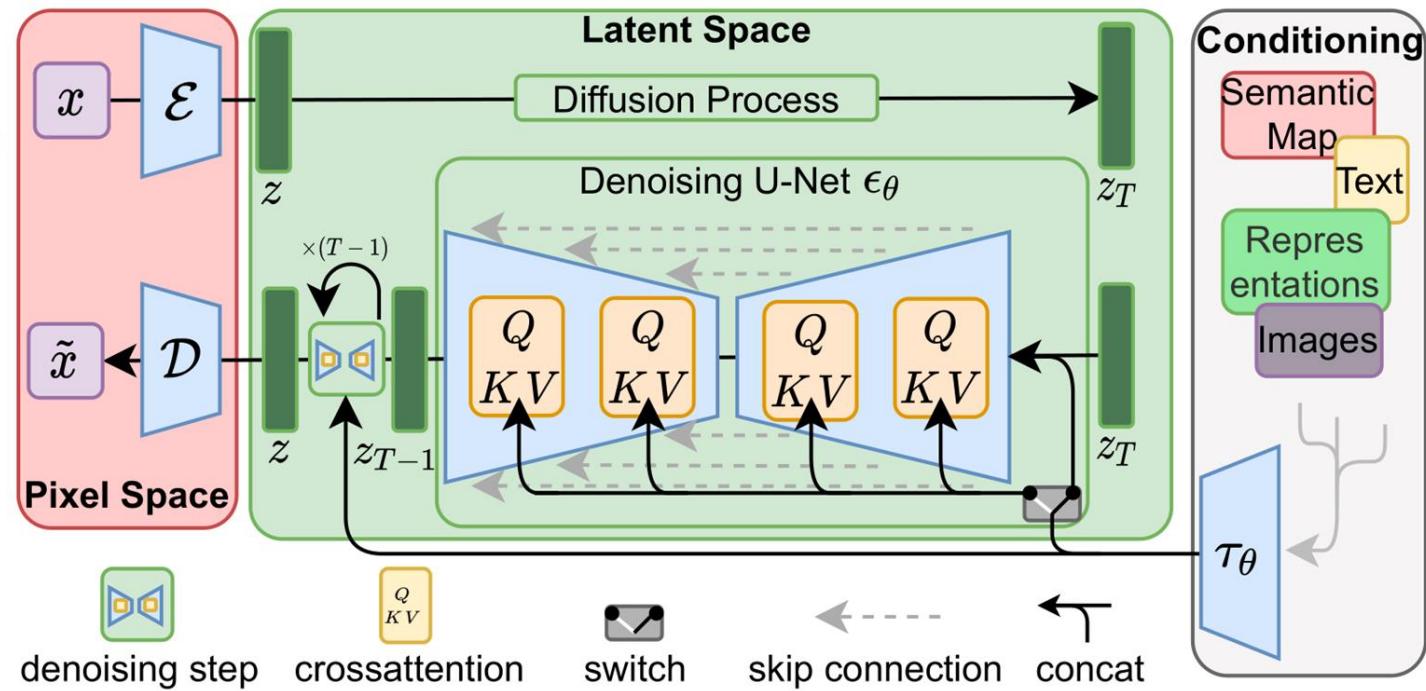
# Stable Diffusion

## Text Encoder - CLIP



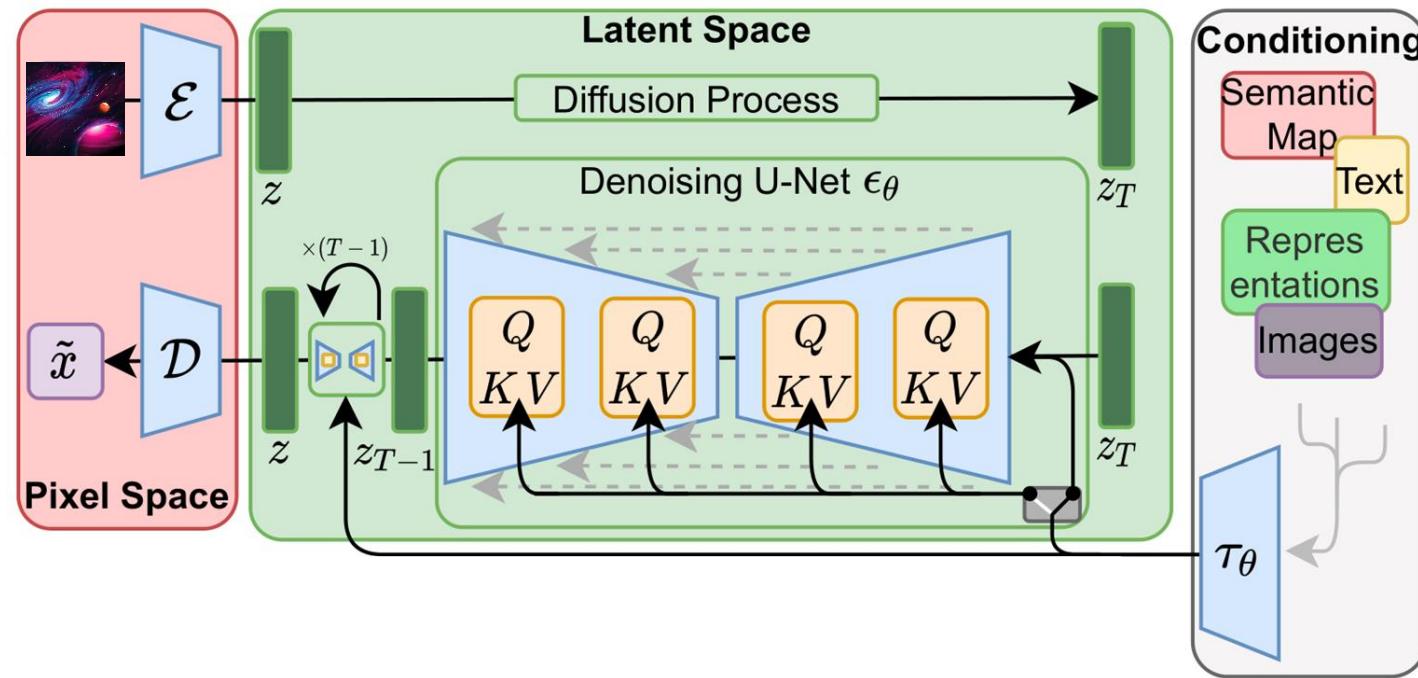
# Stable Diffusion

## Treinamento



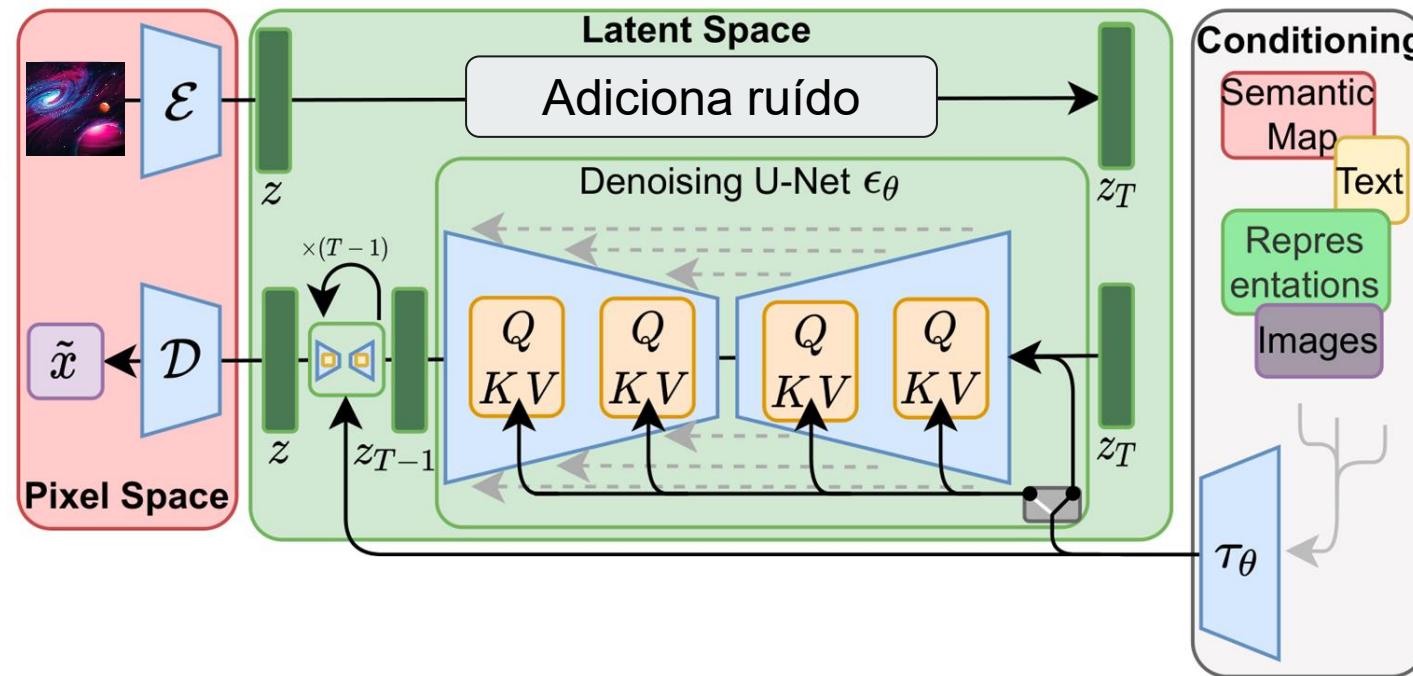
# Stable Diffusion

## Treinamento



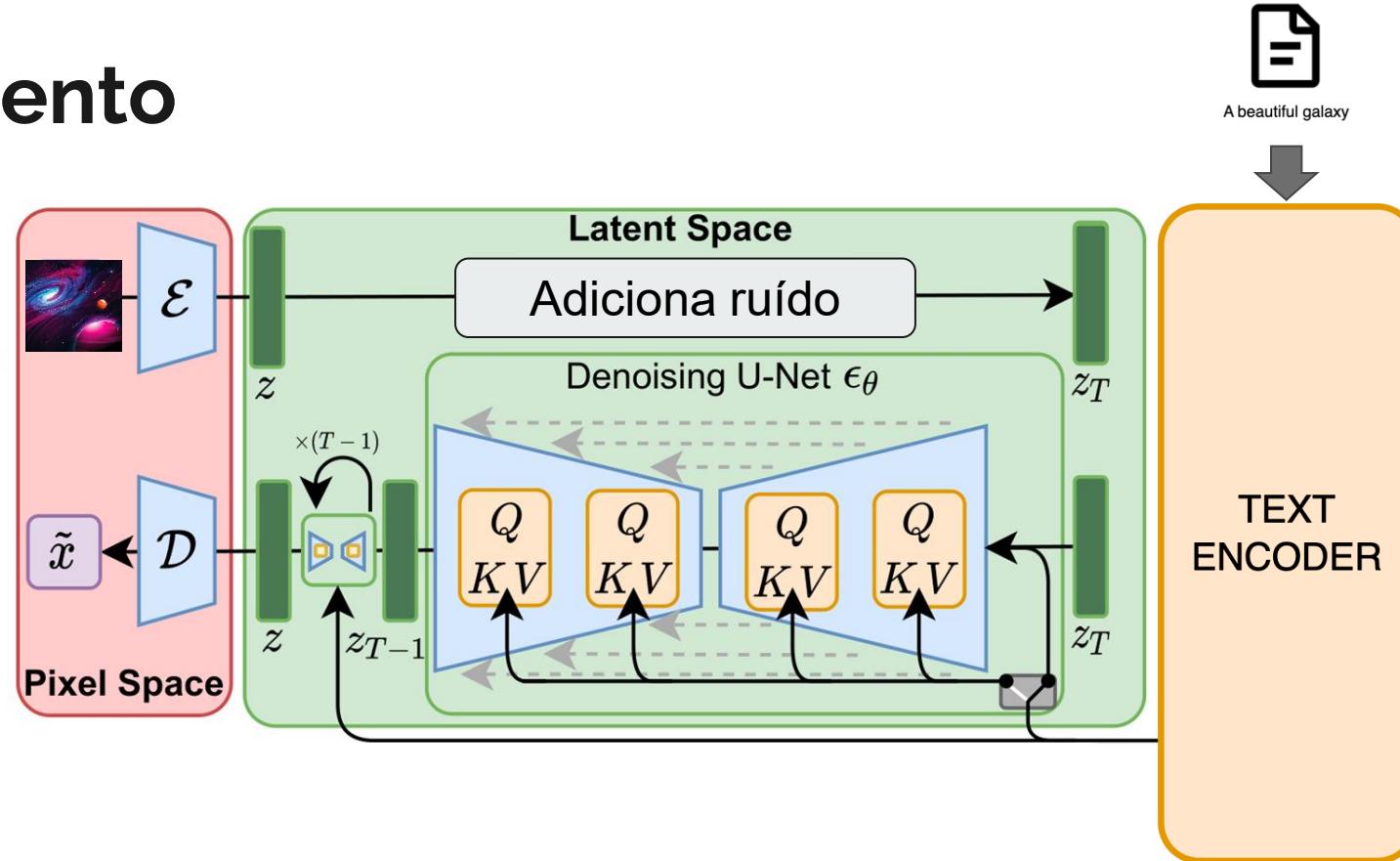
# Stable Diffusion

## Treinamento



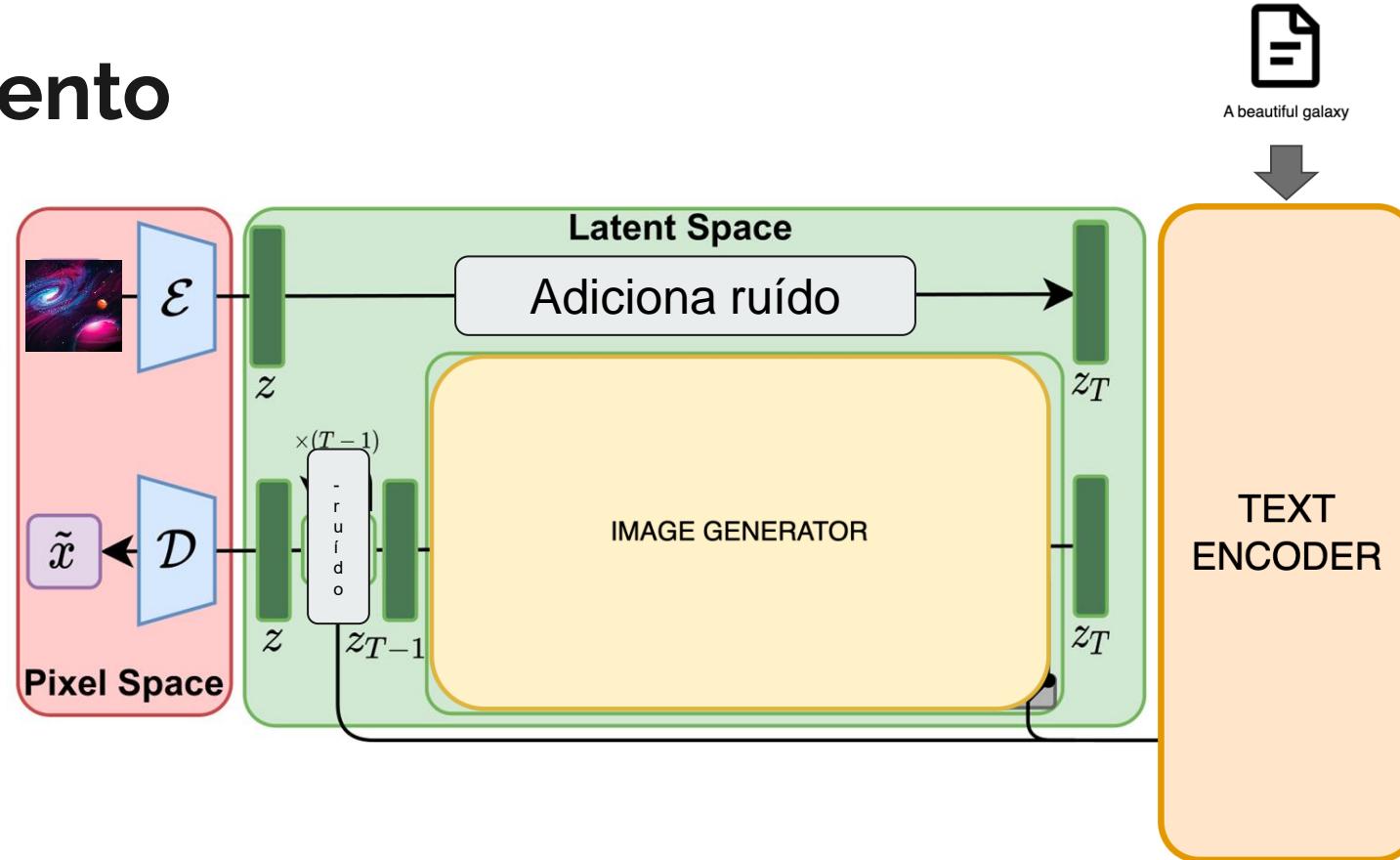
# Stable Diffusion

## Treinamento



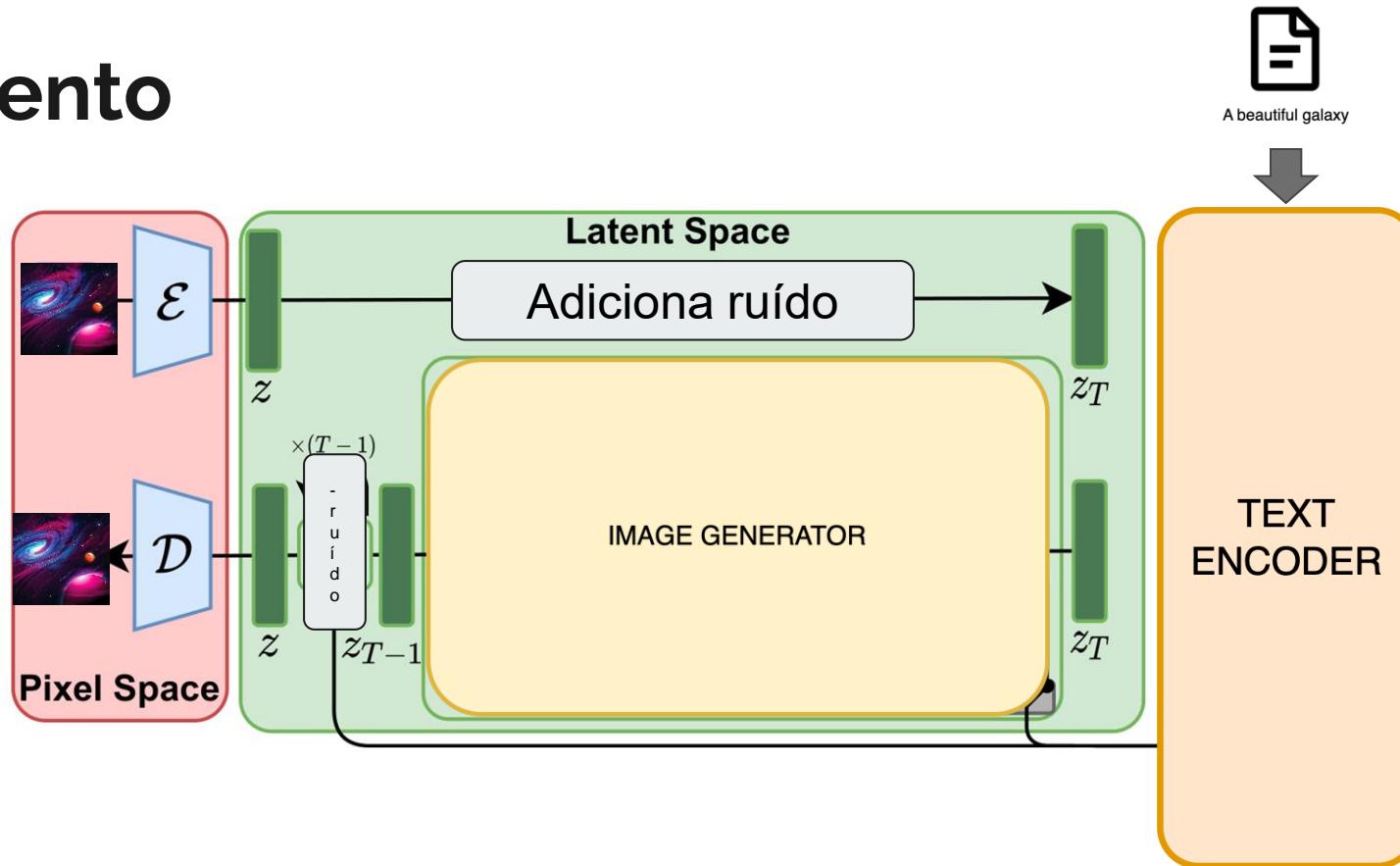
# Stable Diffusion

## Treinamento



# Stable Diffusion

## Treinamento



# Textual Inversion

## 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 ✓



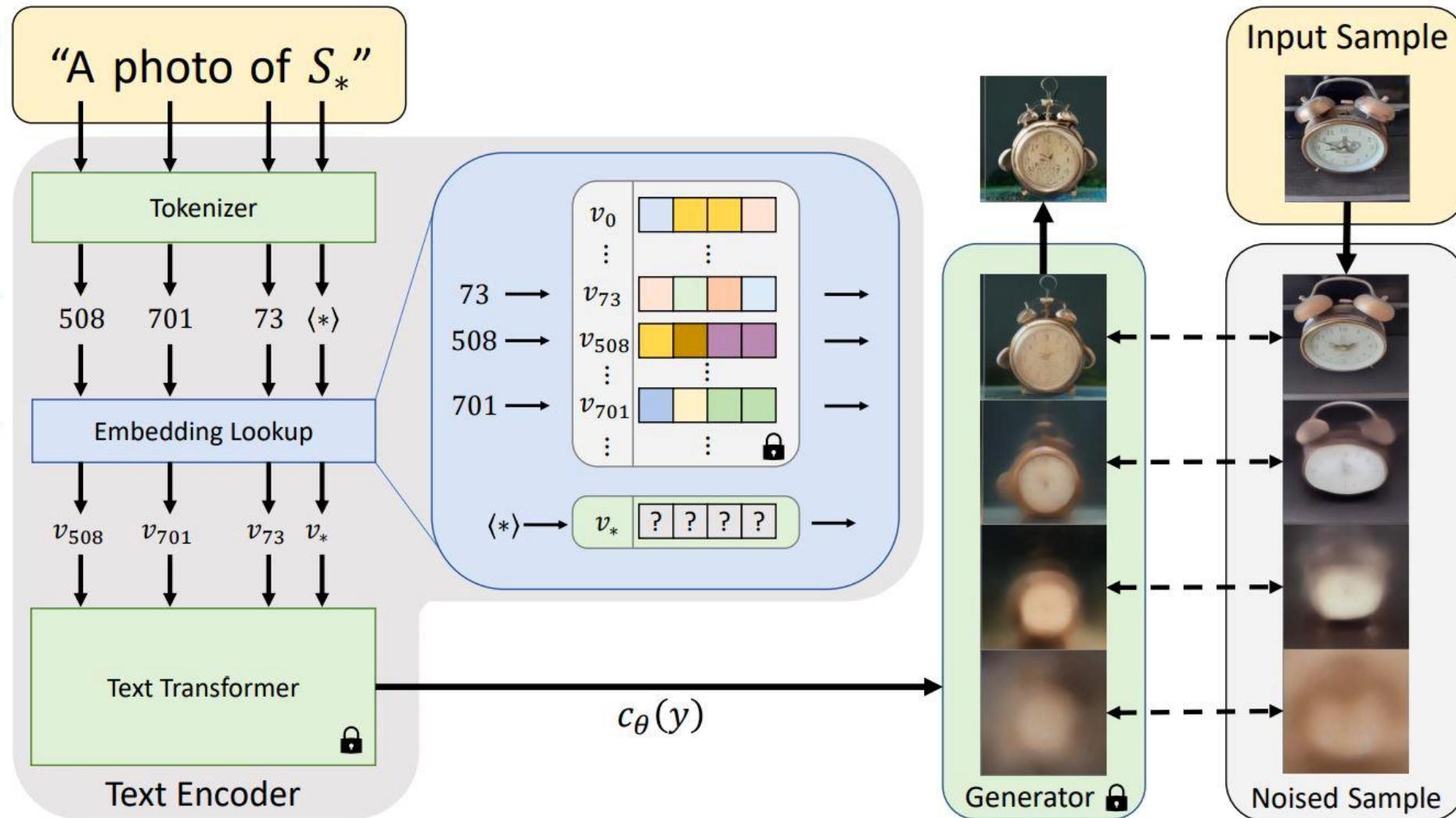
# Textual Inversion

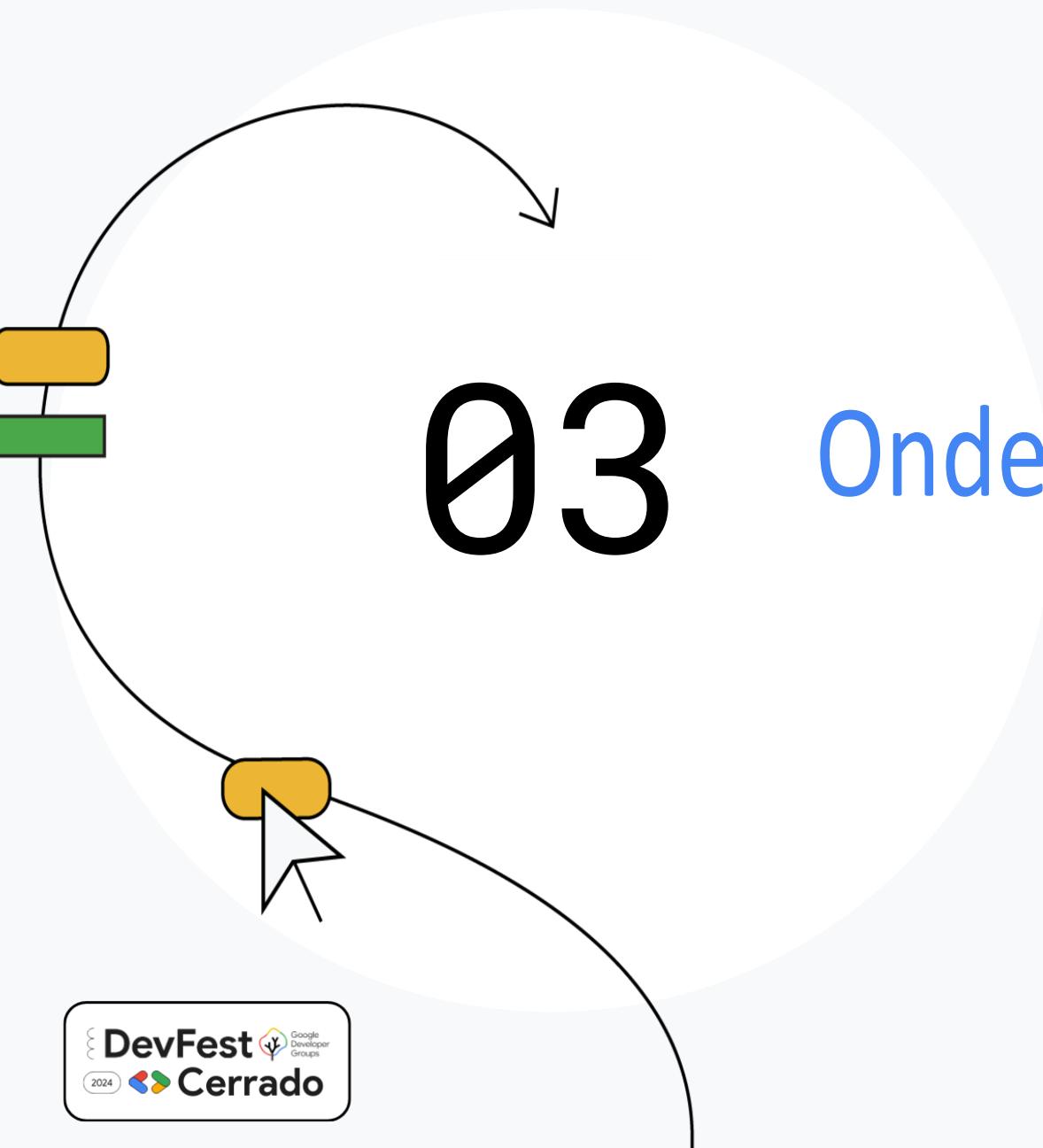
## Dados fora do treinamento

- Um gato em uma praça ✓
- Um cachorro na grama ✓
- Um coelho no mato ✗



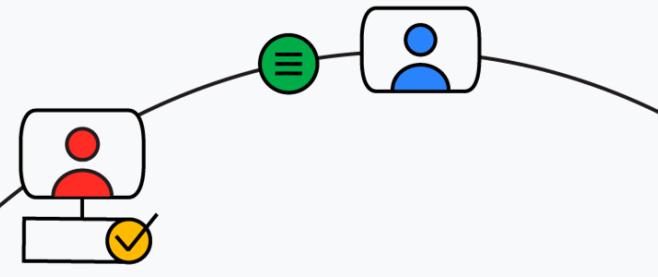
# Textual Inversion



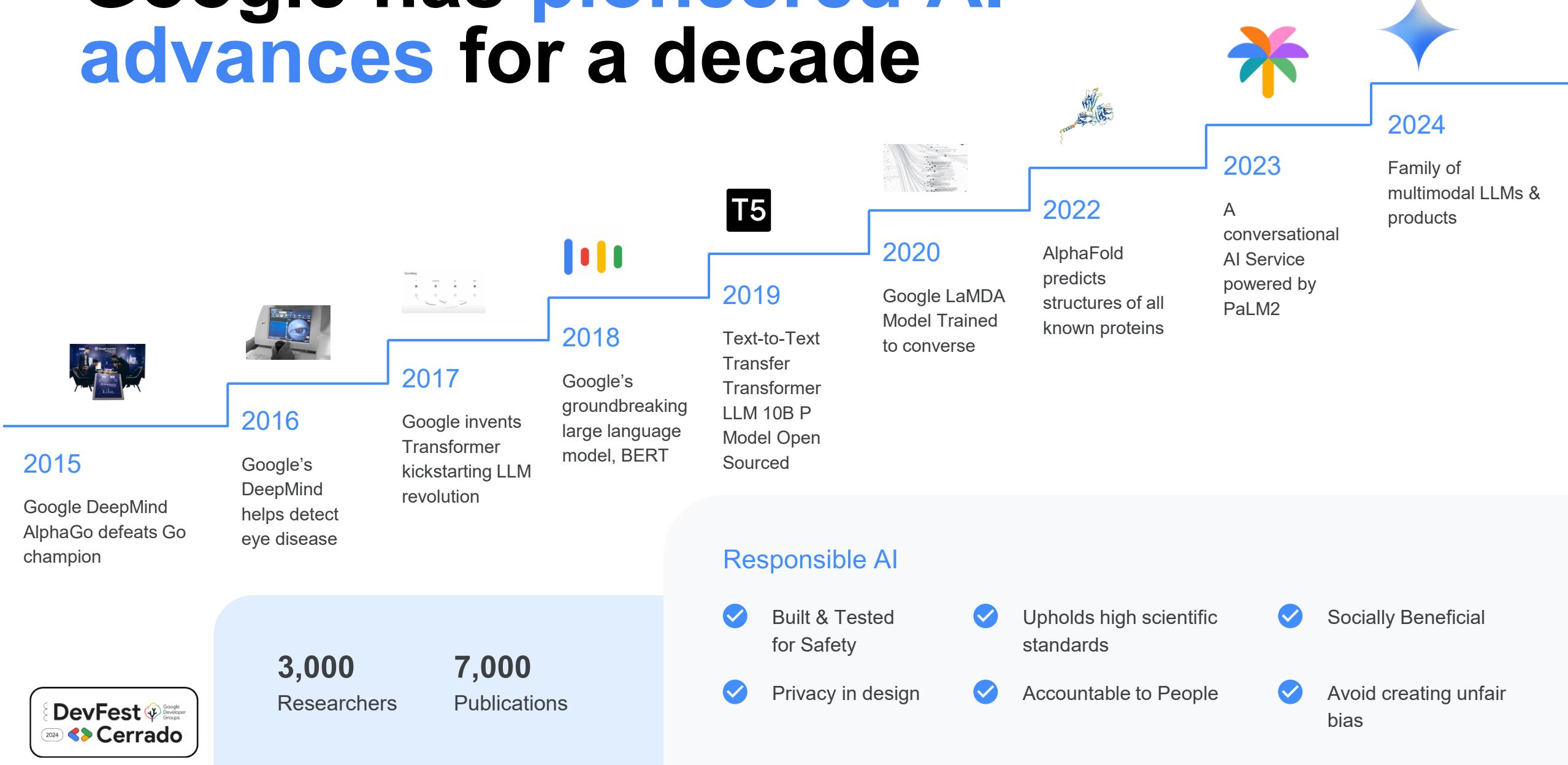


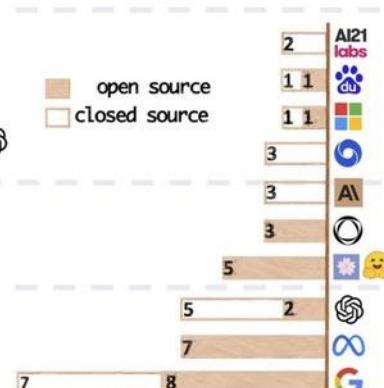
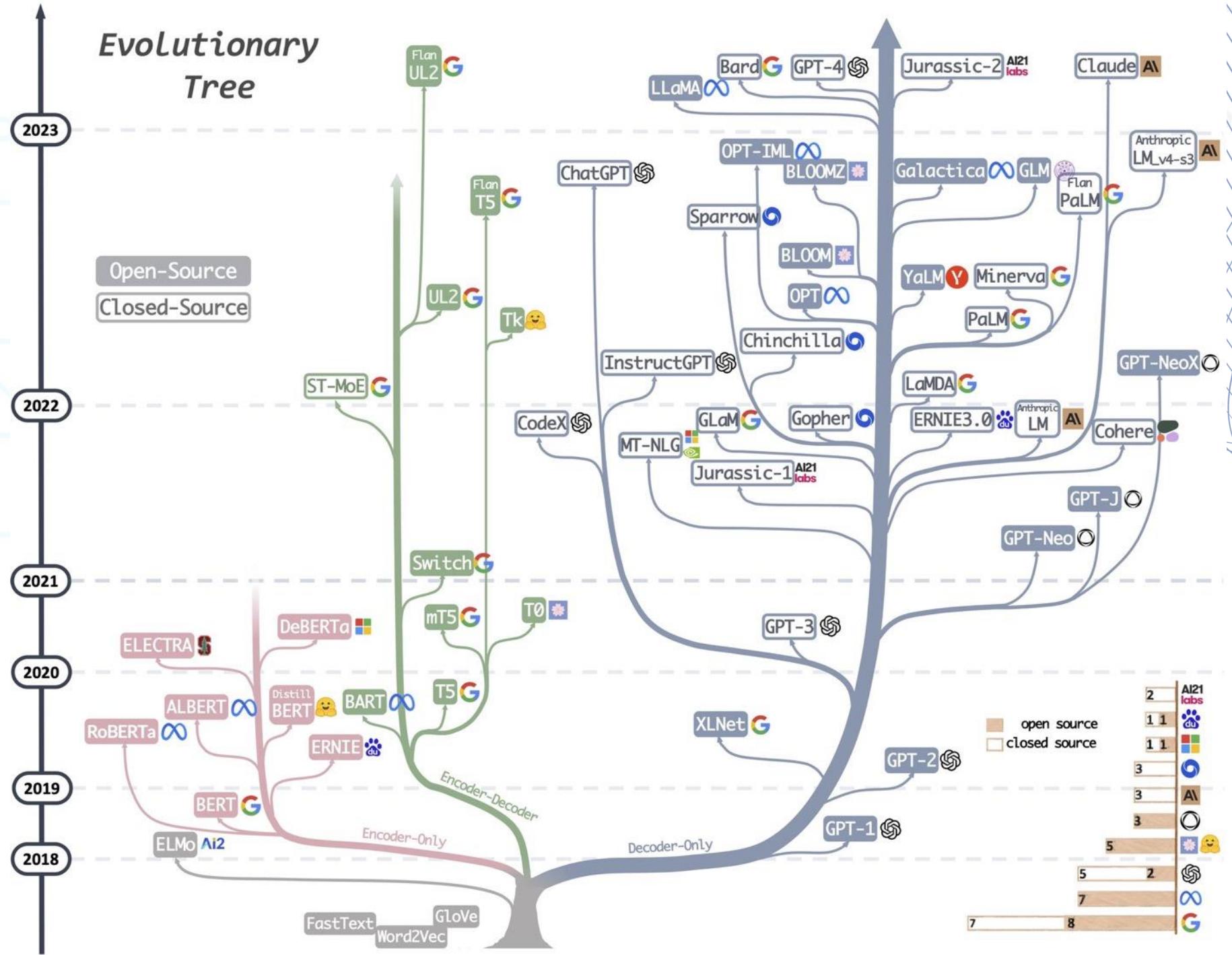
03

Onde o Google entra nisso?

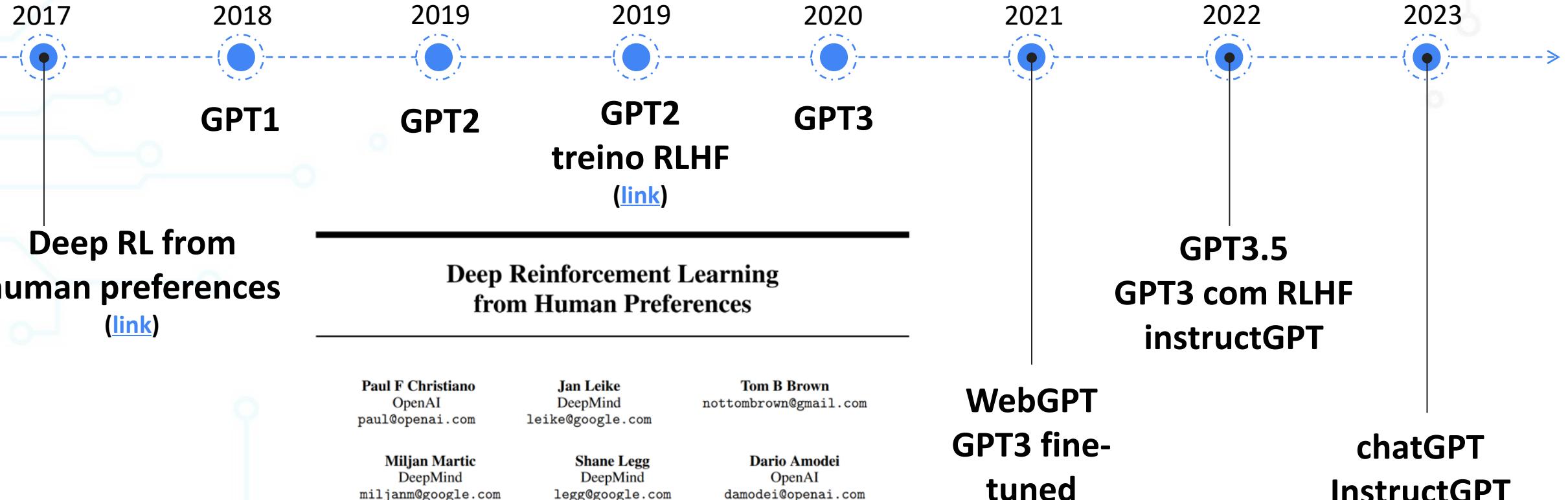


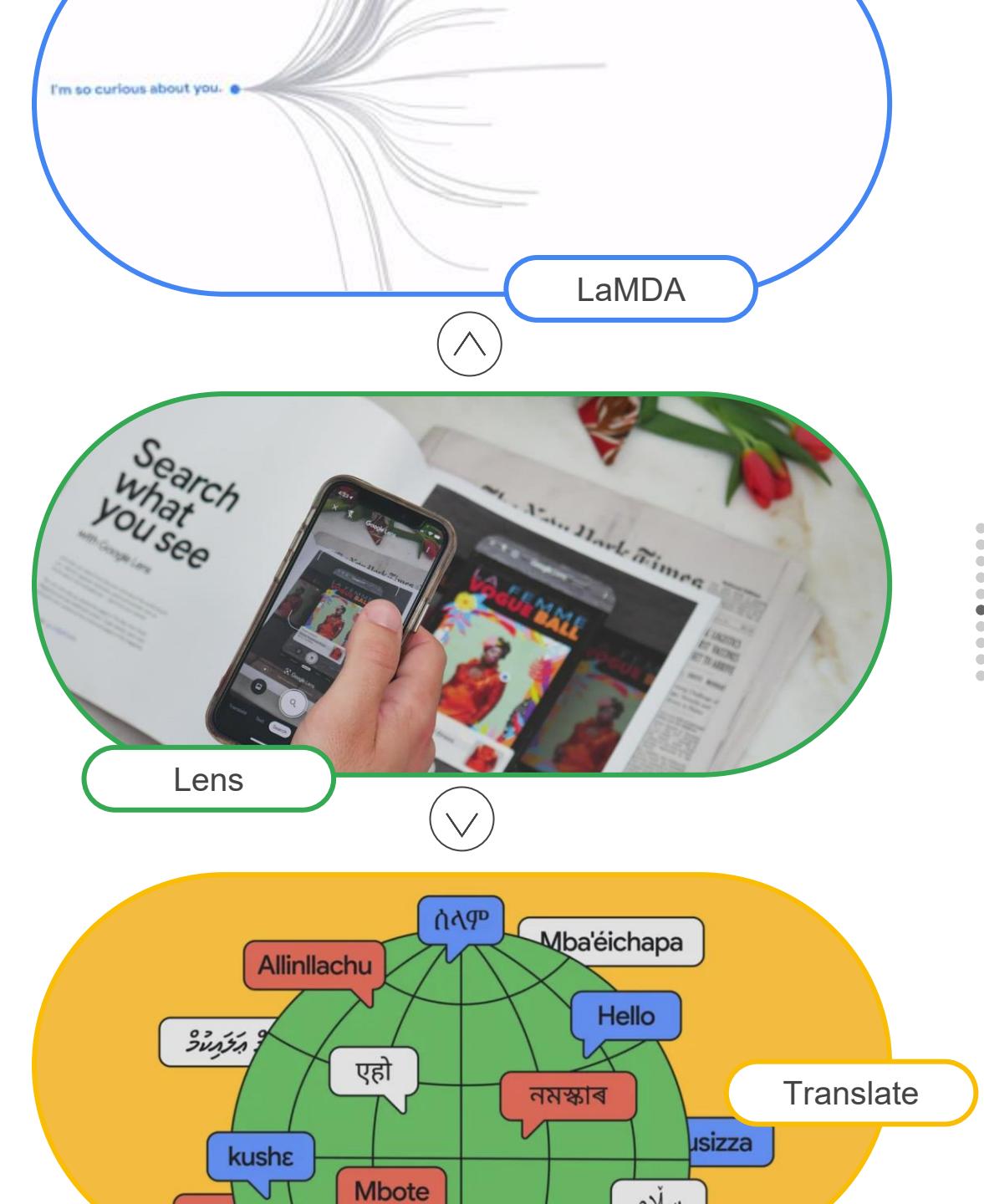
# Google has pioneered AI advances for a decade





# Linha do tempo até o chat Generative Pre-Training chatGPT





Ultimately  
helping  
**people**  
achieve their  
potential.

# Google's AI Principles



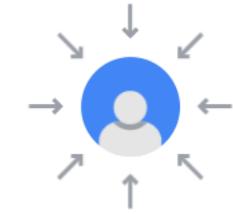
1. Be socially beneficial.



2. Avoid creating or reinforcing unfair bias.



3. Be built and tested for safety.



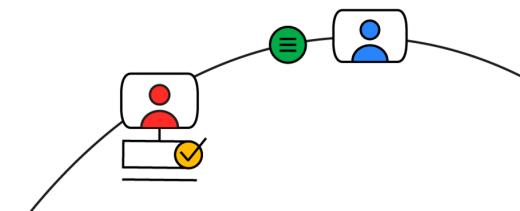
4. Be accountable to people.



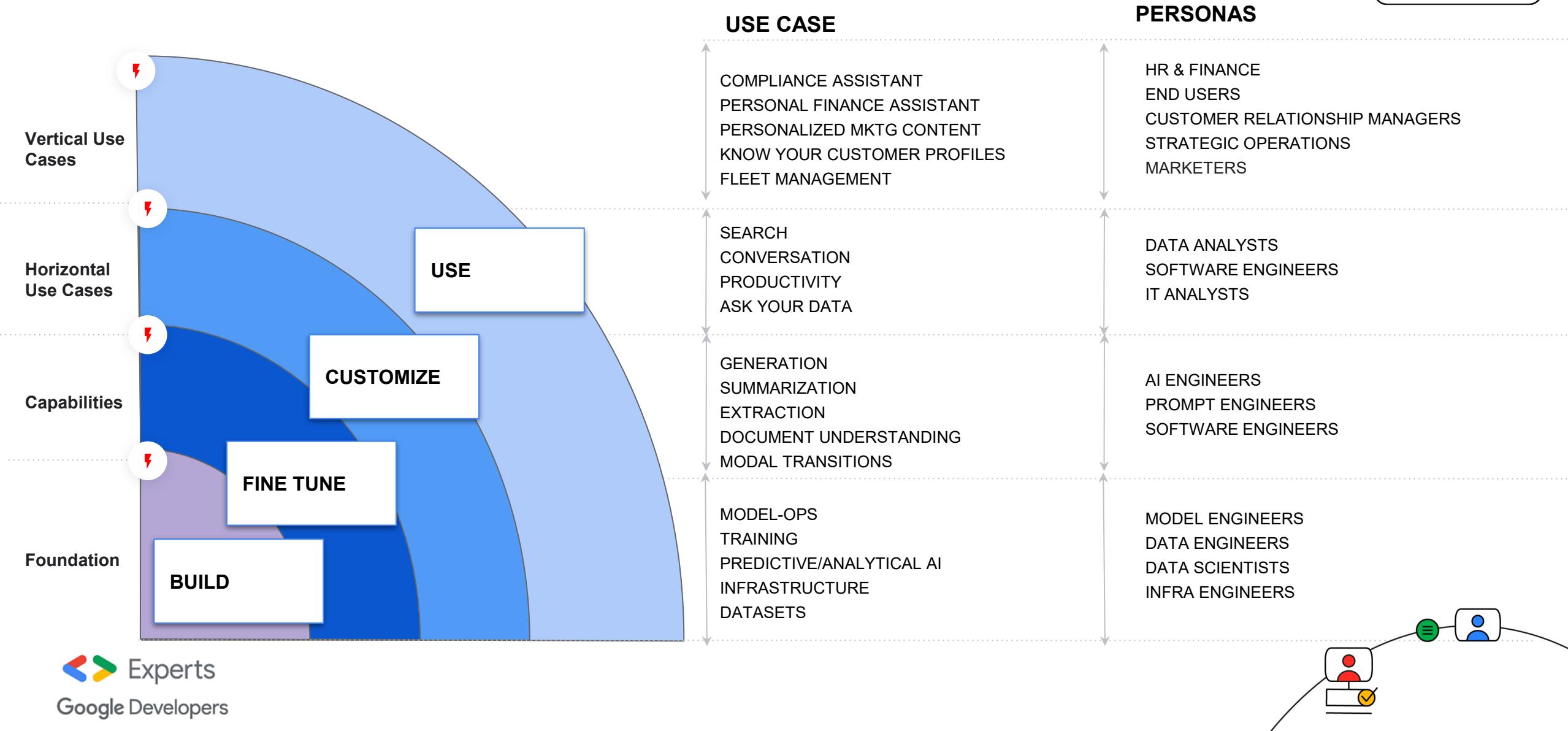
5. Incorporate privacy design principles.

6. Uphold high standards of scientific excellence.

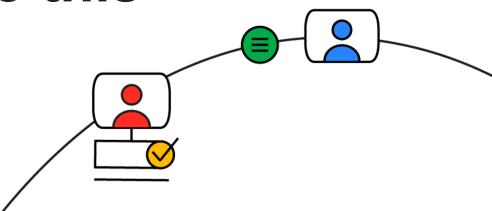
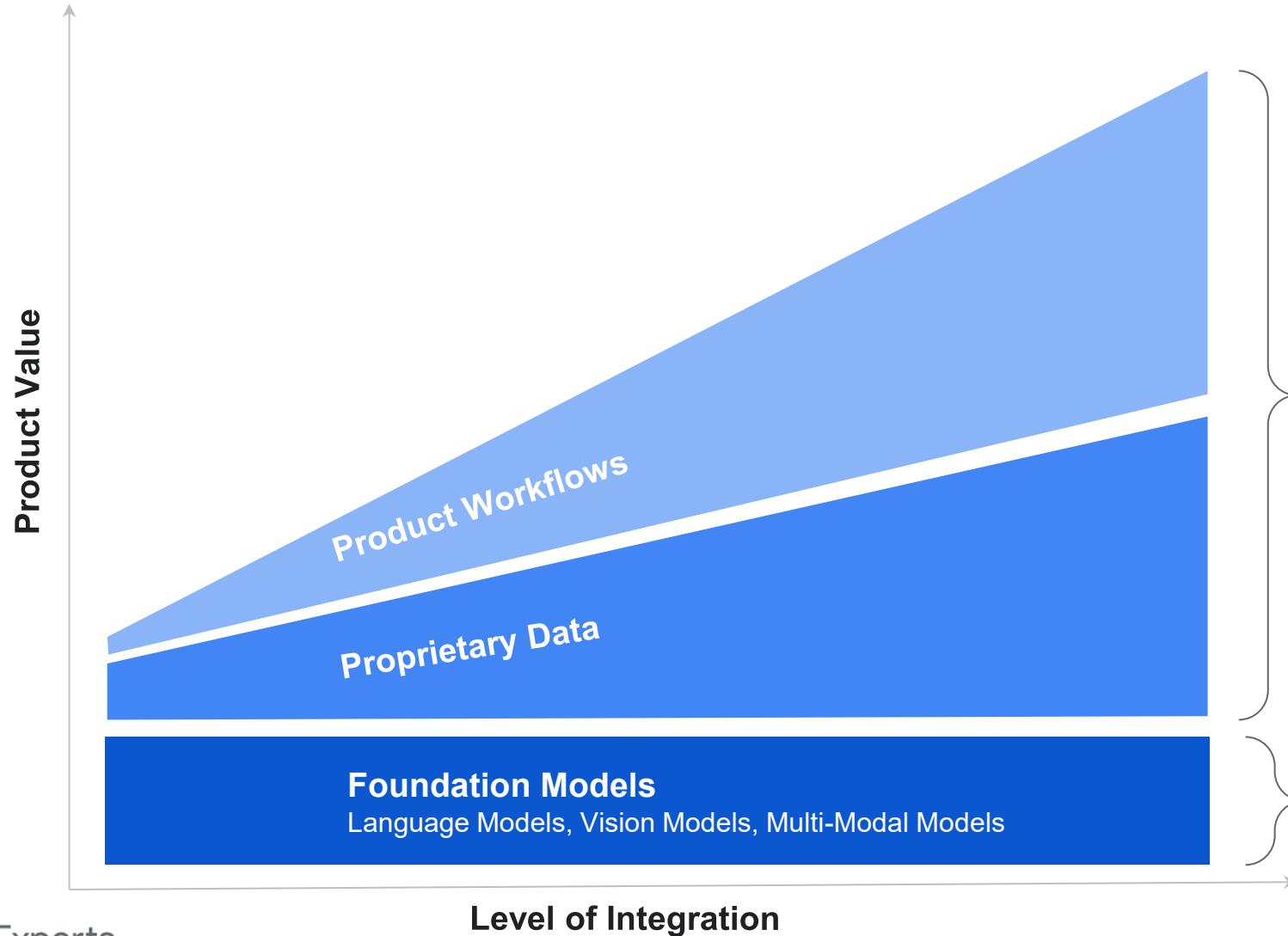
7. Be made available for uses that accord with these principles.



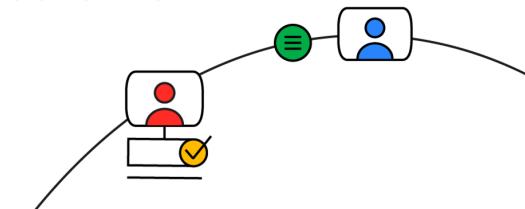
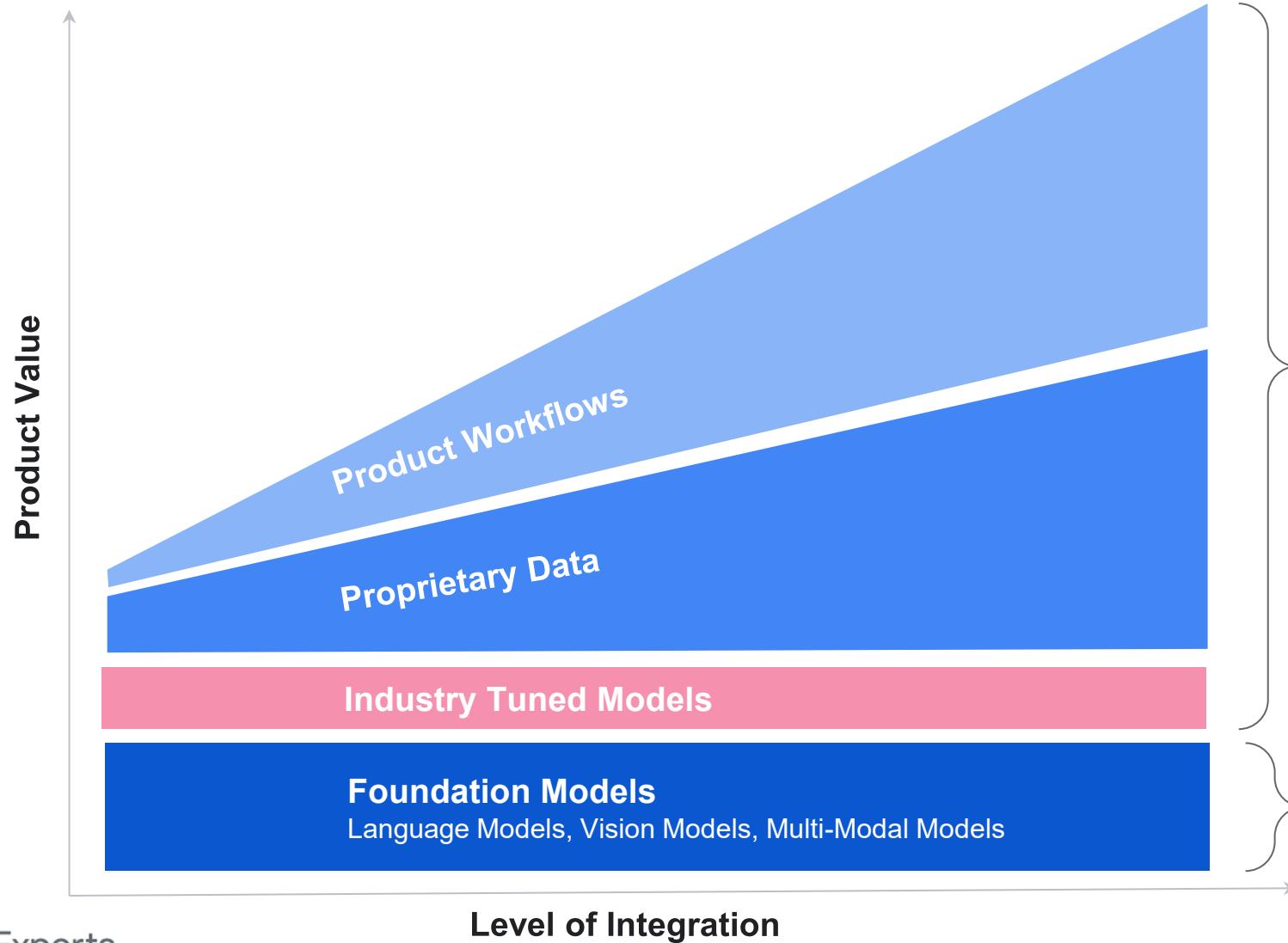
# The Layer Cake of GenAI



# Where Value Accrues with GenAI



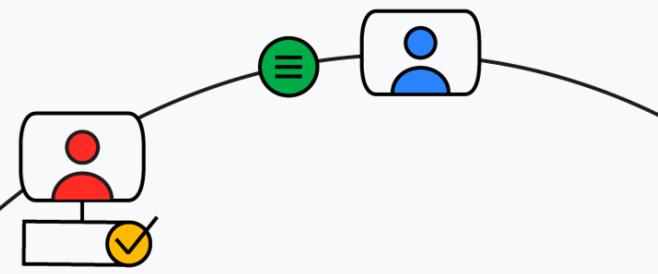
# Where Value Accrues with GenAI





04

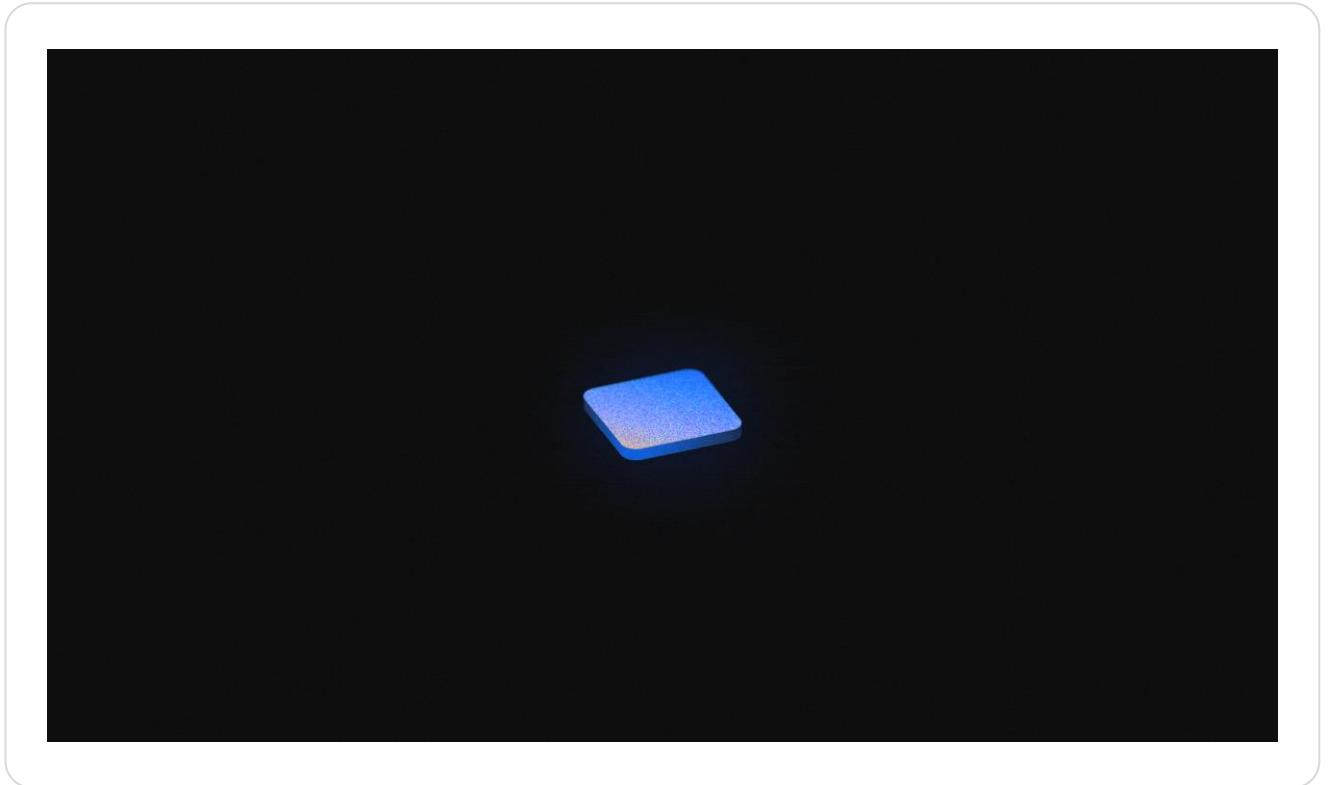
# The Gemini and Gemma Era



# The next chapter of Generative AI innovation



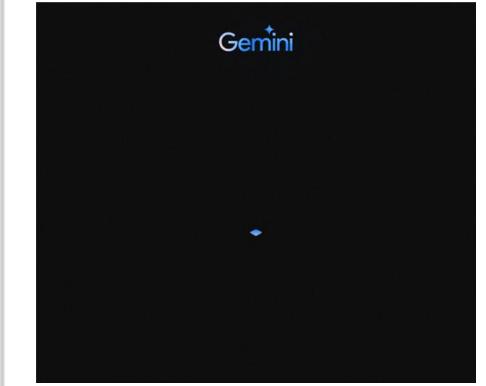
Gemini is the most capable and general model we've ever built, and is the result of a large-scale collaborative effort by teams across Google, including Google DeepMind and Google Research.



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



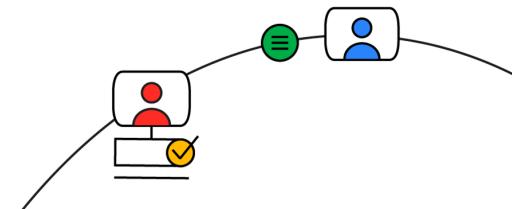
State-of-the-art, natively  
multimodal reasoning  
capabilities



Highly optimized while  
preserving choice

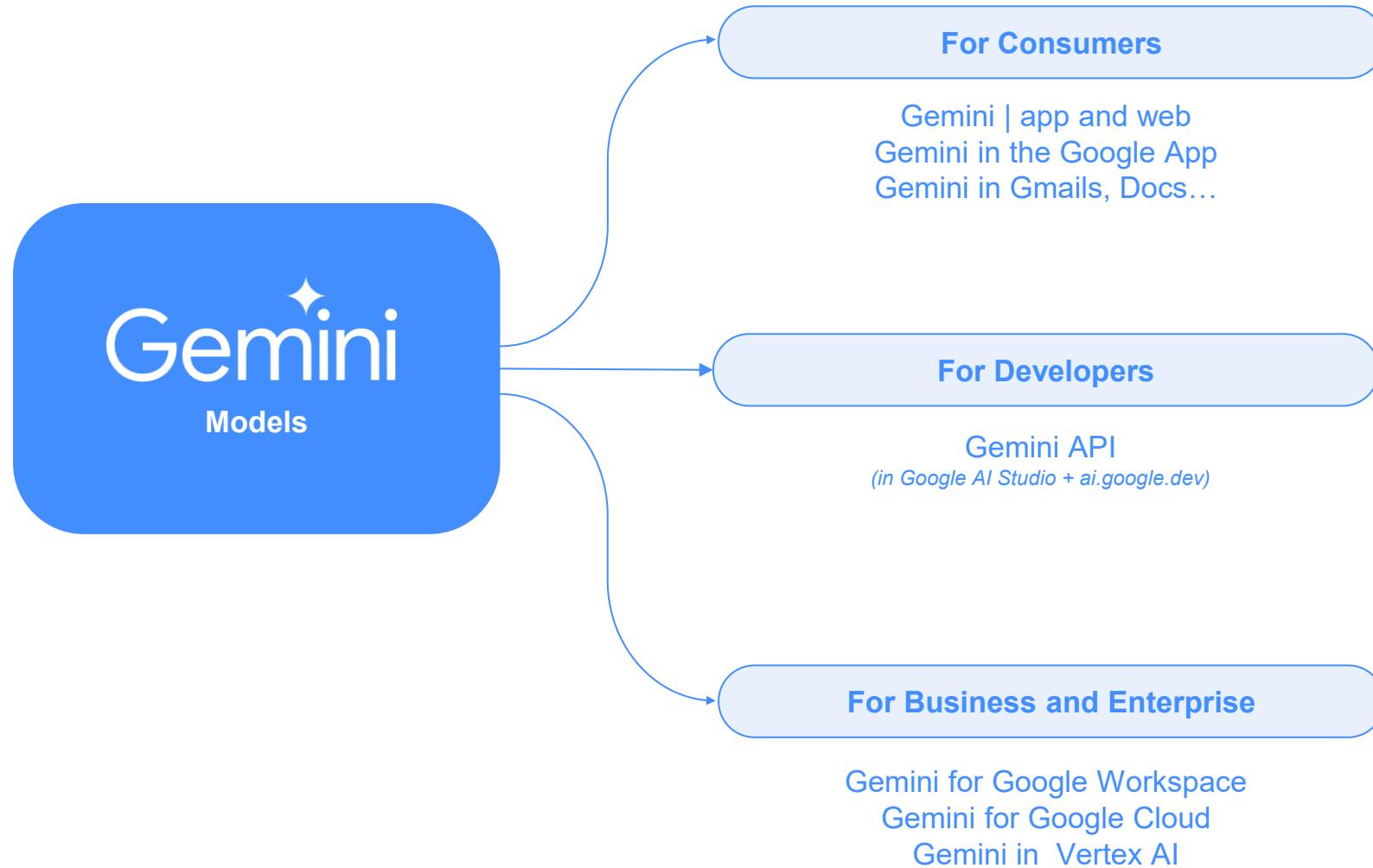


Built with responsibility  
and safety at the core

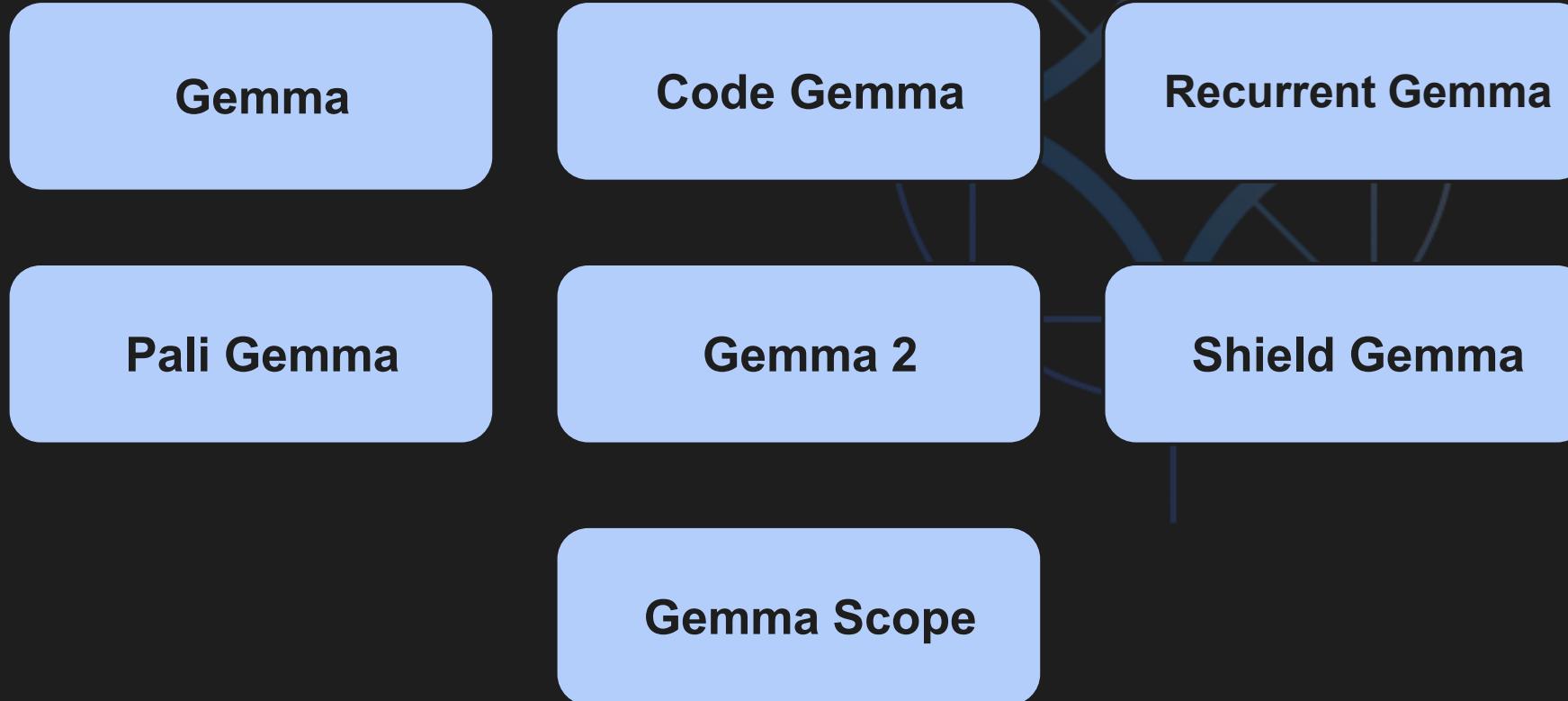


# The Gemini Ecosystem

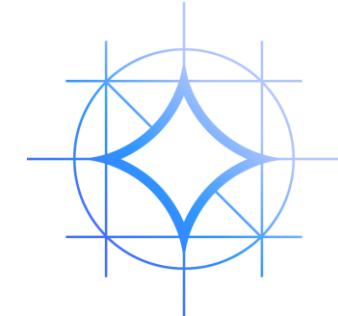
The most advanced AI from Google



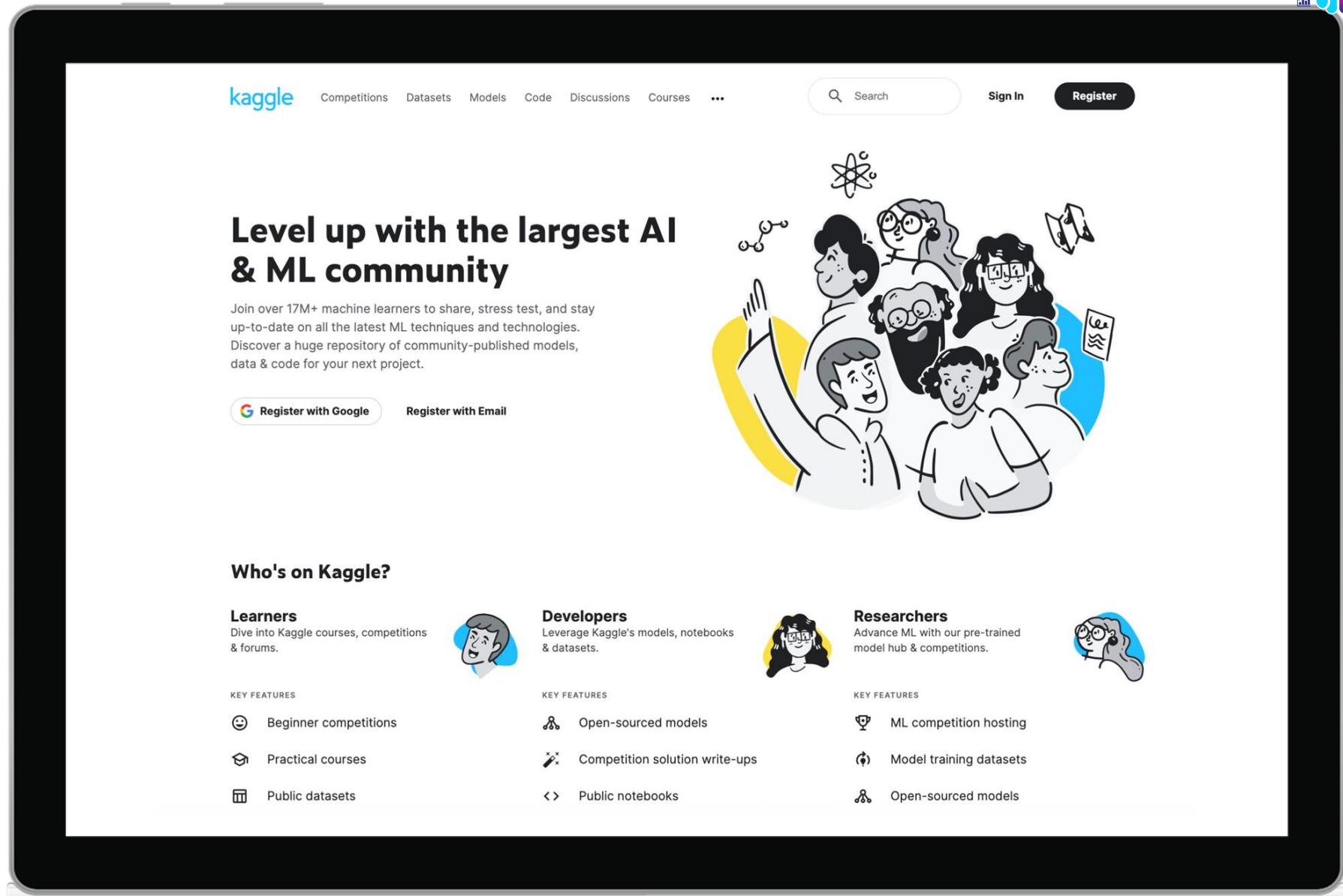
# Gemma Family



kaggle



Gemma



The image shows a tablet displaying the Kaggle homepage. The top navigation bar includes links for Competitions, Datasets, Models, Code, Discussions, Courses, and a search bar. On the right side of the header are 'Sign In' and 'Register' buttons. The main headline reads 'Level up with the largest AI & ML community'. Below it, a sub-headline encourages users to join over 17M+ machine learners. A large, colorful illustration of a diverse group of people is centered on the page. At the bottom, there's a section titled 'Who's on Kaggle?' with three categories: 'Learners', 'Developers', and 'Researchers', each with a brief description and associated icons.

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

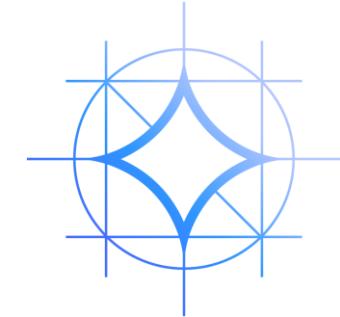
Advance ML with our pre-trained model hub & competitions.

KEY FEATURES

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



**K** Keras



# Gemma

# Why Keras?



Simple

simplifies development,  
empowering focused  
problem-solving.



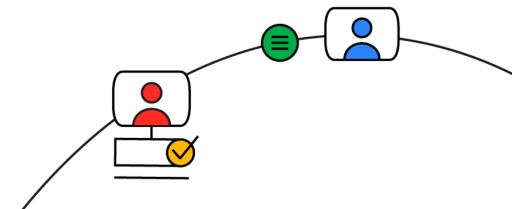
Flexible

offers scalable complexity  
for evolving needs.



Powerful

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



Model Garden  
Google | OSS | Partner Models



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

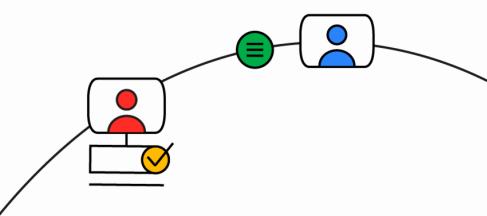


Google Developers

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**.



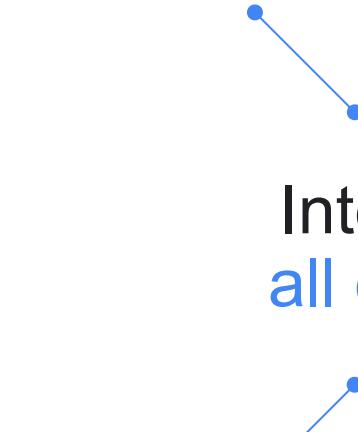
**Packages and extensions** to natively support Google Cloud foundation models in Google app developer frameworks like **Firebase** and **Flutter**.



Vertex AI



Colab



Interfaces for  
all developers



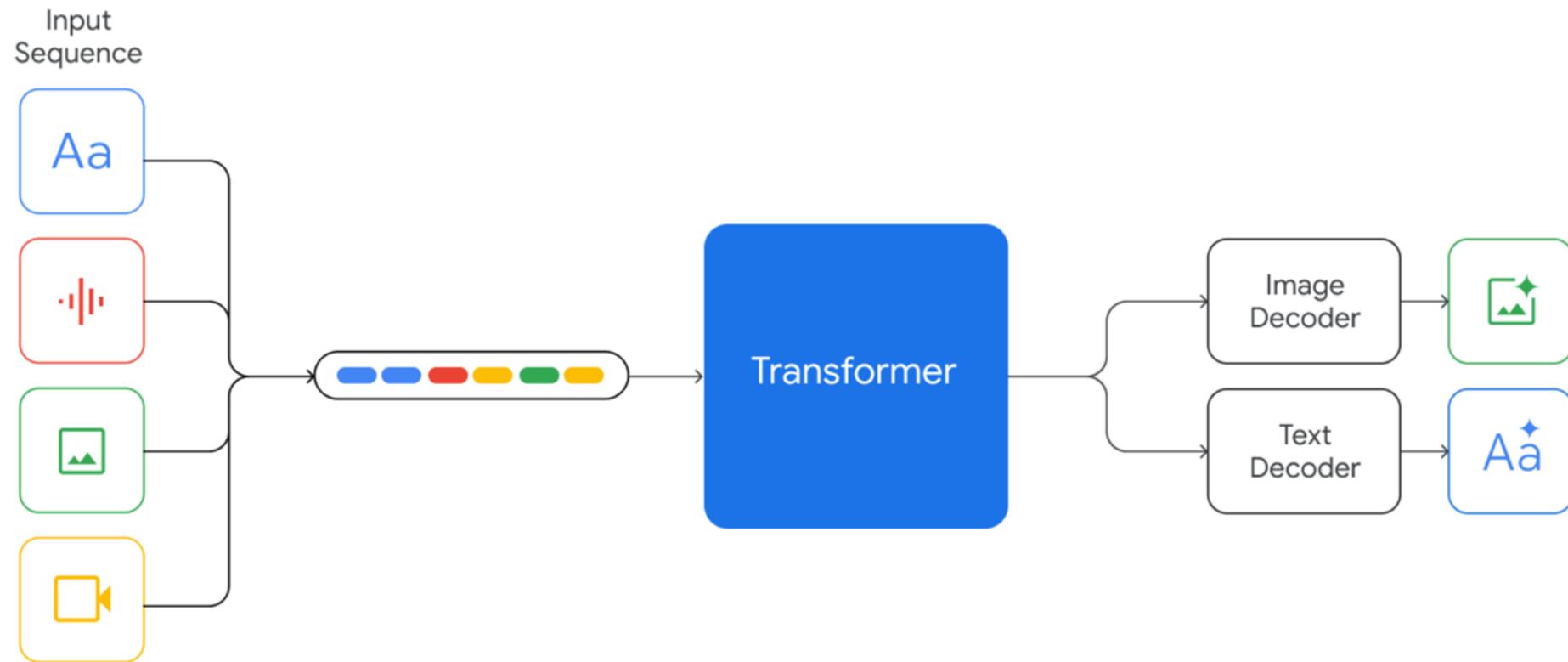
Flutter



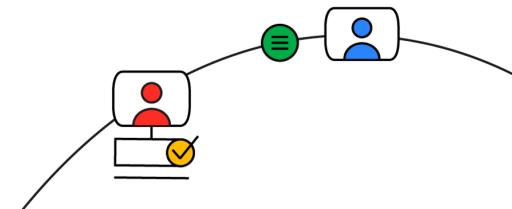
Firebase



# Multimodality



# Gemini vs Gemma



# Purpose

## Gemini

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

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.

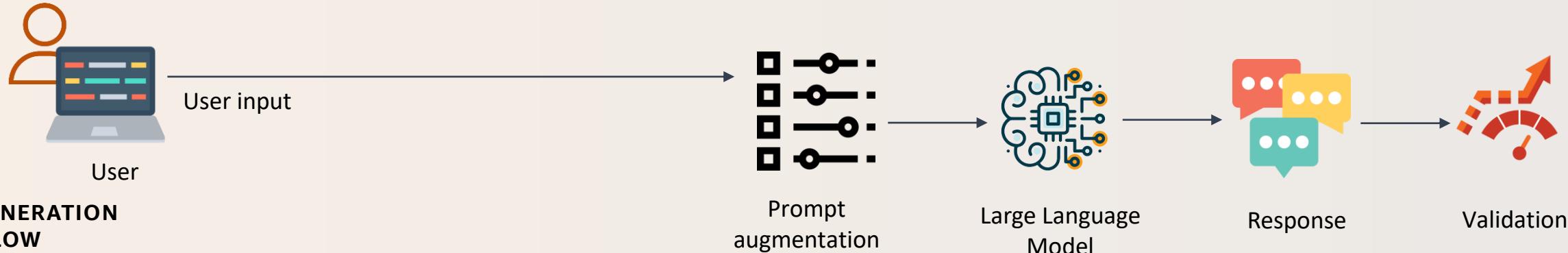
# Integration

## Gemini

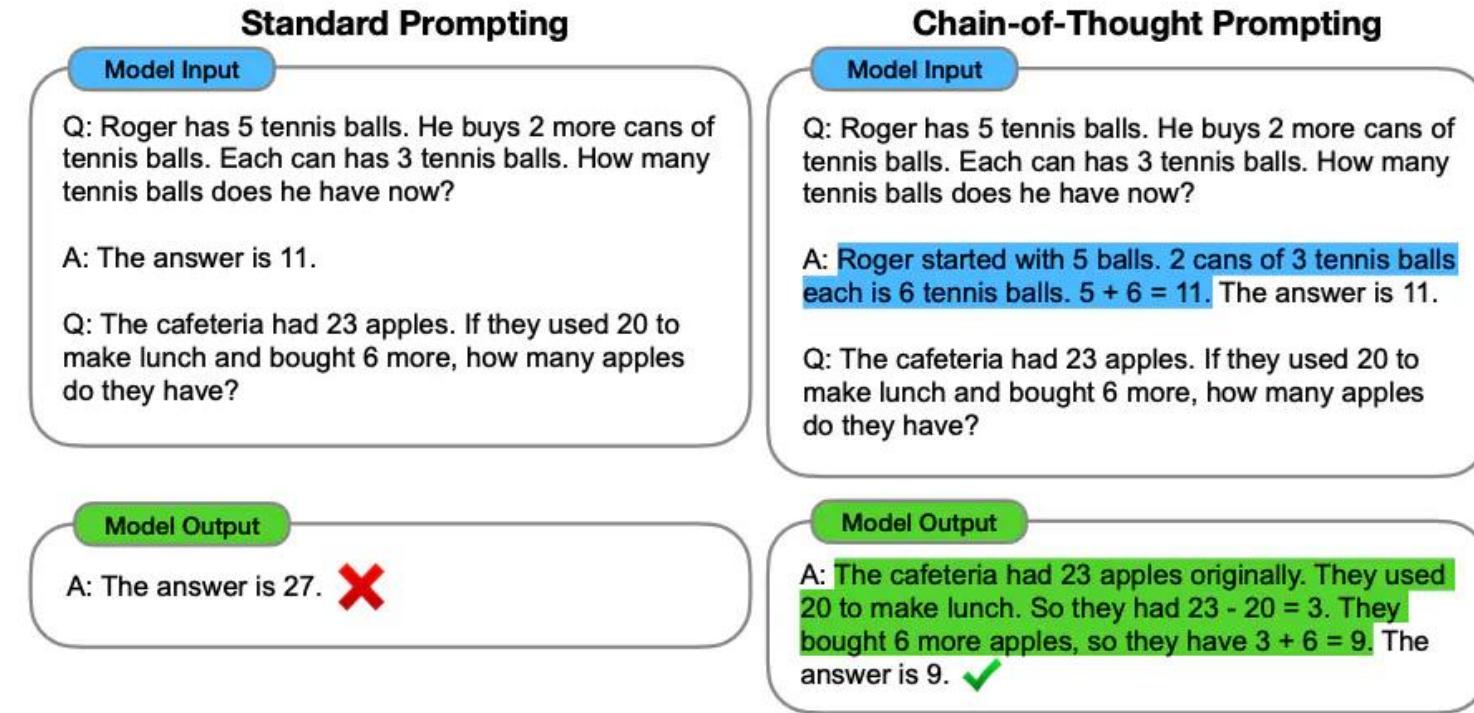
Gemini is integrated into Google services like Bard and Search, and can be used via Google Cloud for NLP-heavy applications.

## Gemma

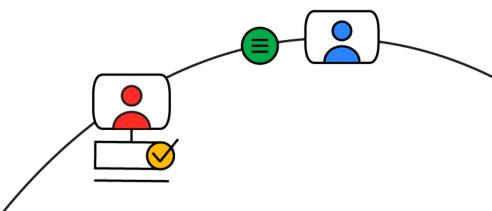
It's also deeply integrated into Google Cloud's Vertex AI ecosystem, making it easy to use for developers and data scientists who are already working within the Google Cloud environment.



# Chain-of-Thought Prompting (CoT)



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



# Self Consistency



## Self-consistency

**Q:** If there are 3 cars in the parking lot and 2 more cars arrive, how many cars are in the parking lot?

**A:** There are 3 cars in the parking lot already. 2 more arrive. Now there are  $3 + 2 = 5$  cars. The answer is 5.

...  
**Q:** Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?

**A:**

Language model

Sample a diverse set of reasoning paths

Marginalize out reasoning paths to aggregate final answers

She has  $16 - 3 - 4 = 9$  eggs left. So she makes  $\$2 * 9 = \$18$  per day.

The answer is \$18.

This means she sells the remainder for  $\$2 * (16 - 4 - 3) = \$26$  per day.

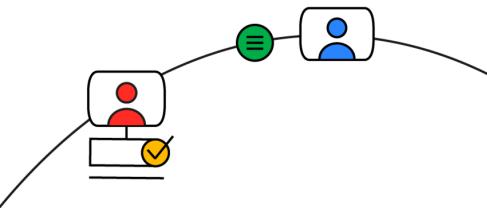
The answer is \$26.

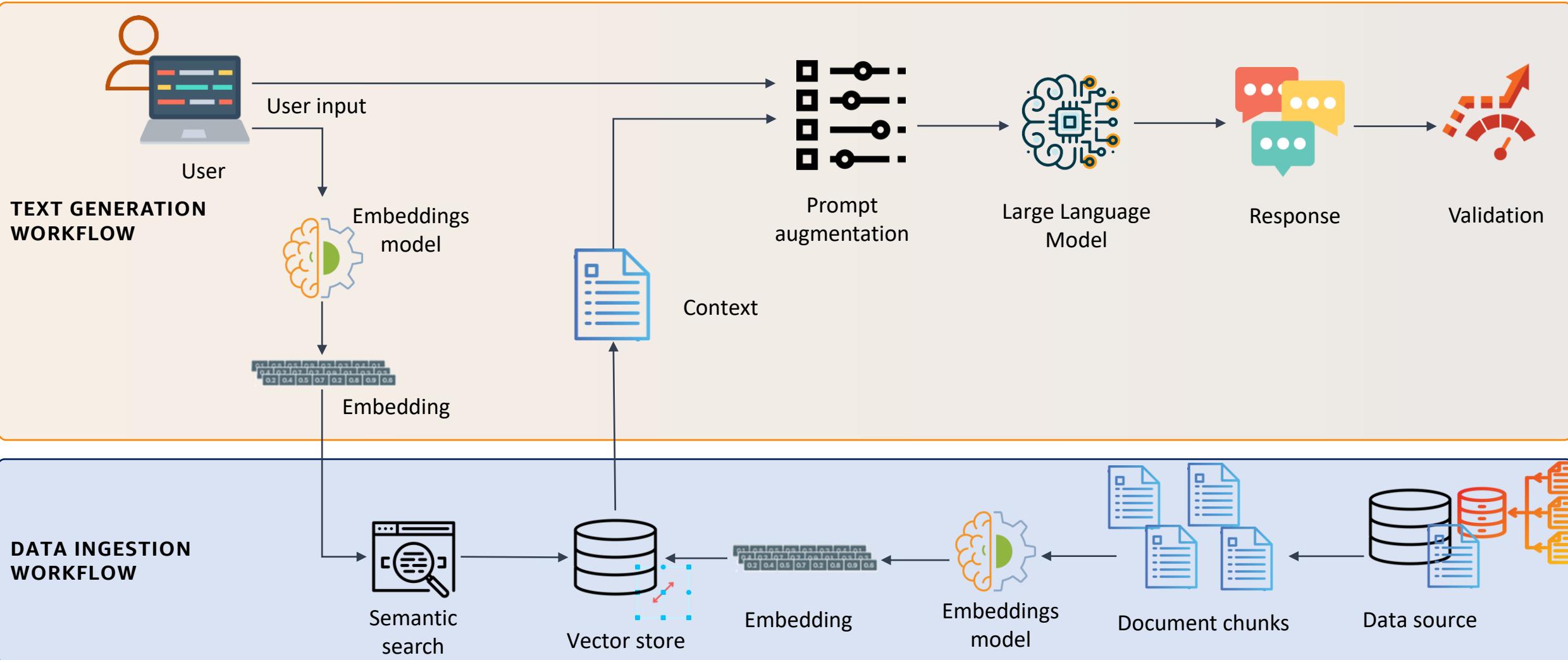
She eats 3 for breakfast, so she has  $16 - 3 = 13$  left. Then she bakes muffins, so she has  $13 - 4 = 9$  eggs left. So she has  $9 \text{ eggs} * \$2 = \$18$ .

The answer is \$18.

The answer is \$18.

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





# Retrieval-Augmented Generation for Large Language Models: A Survey

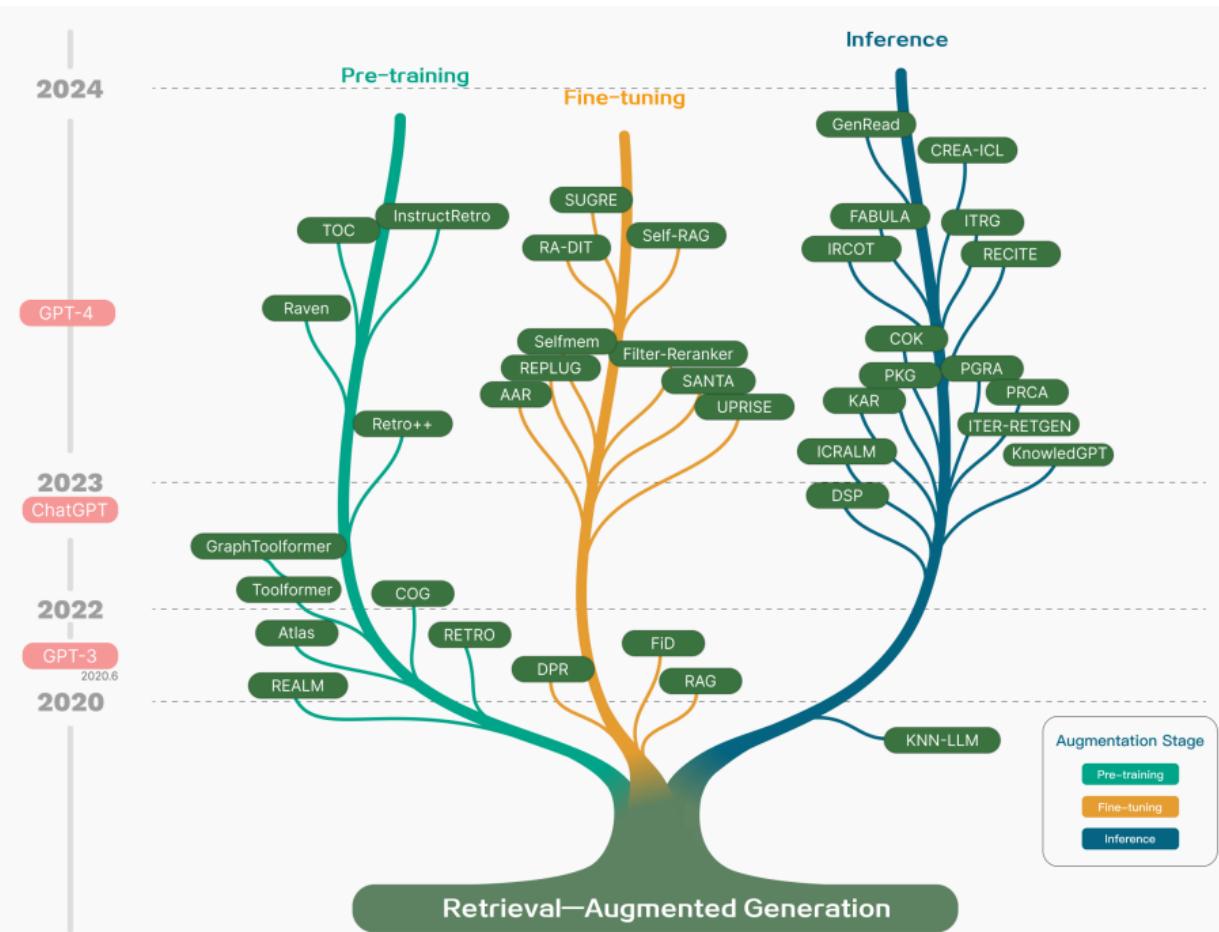
Yunfan Gao<sup>1</sup>, Yun Xiong<sup>2</sup>, Xinyu Gao<sup>2</sup>, Kangxiang Jia<sup>2</sup>, Jinliu Pan<sup>2</sup>, Yuxi Bi<sup>3</sup>, Yi Dai<sup>1</sup>, Jiawei Sun<sup>1</sup>, Qianyu Guo<sup>4</sup>, Meng Wang<sup>3</sup> and Haofen Wang<sup>1,3\*</sup>

<sup>1</sup> Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University

<sup>2</sup> Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University

<sup>3</sup> College of Design and Innovation, Tongji University

<sup>4</sup> 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  
Marcos Piau

Rodrigo Castaldoni  
Pablo Costa

Pedro Gengo  
Vinicius Caridá

Celio Larcher

22h, Brazil

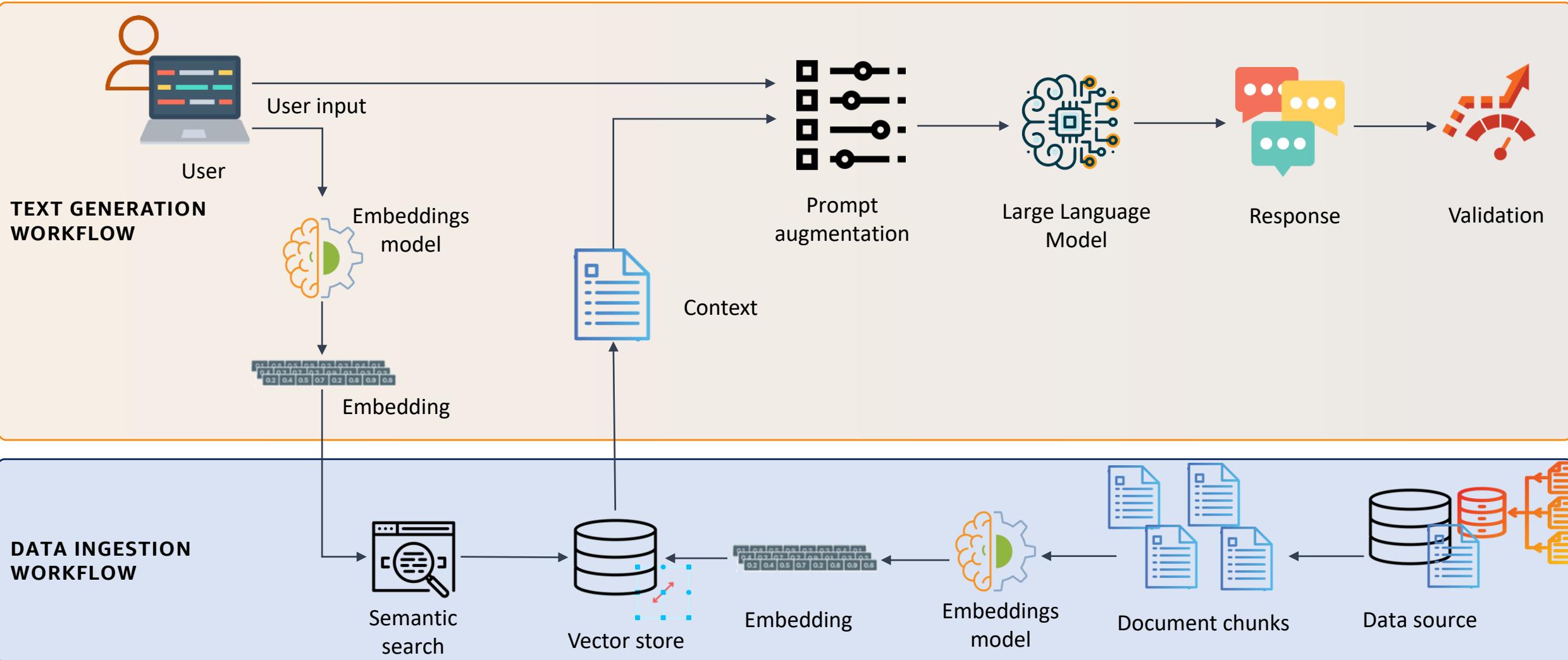
email: {pfinardi, leonardo.bernardi.avila, castaldoniro, pedro.gengo.lourenco, celiolarcher, marcos.piau.vieira, pablo.botton.costa, vfcarida}@gmail.com

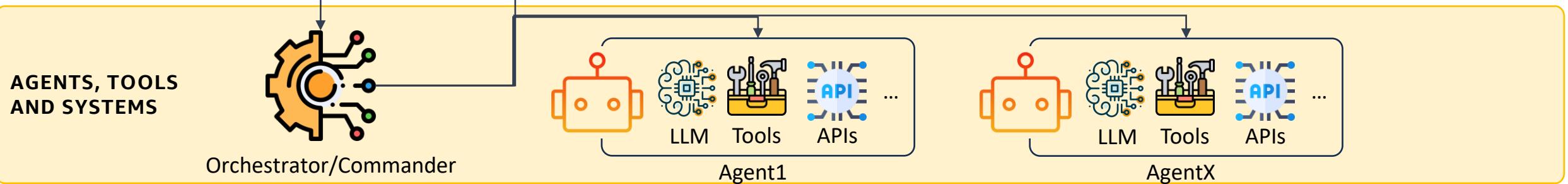
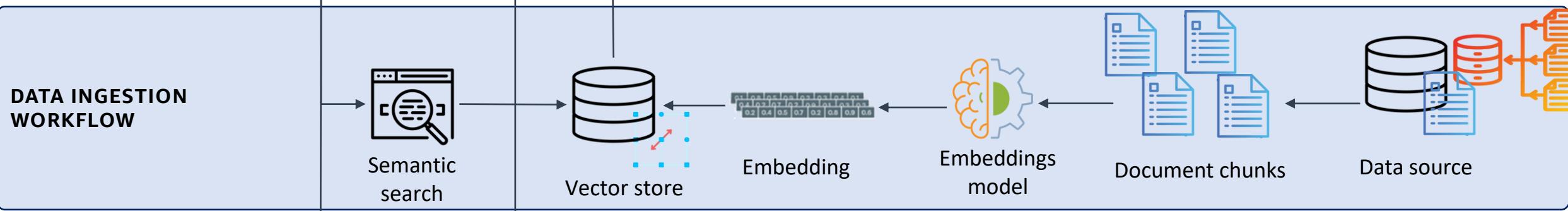
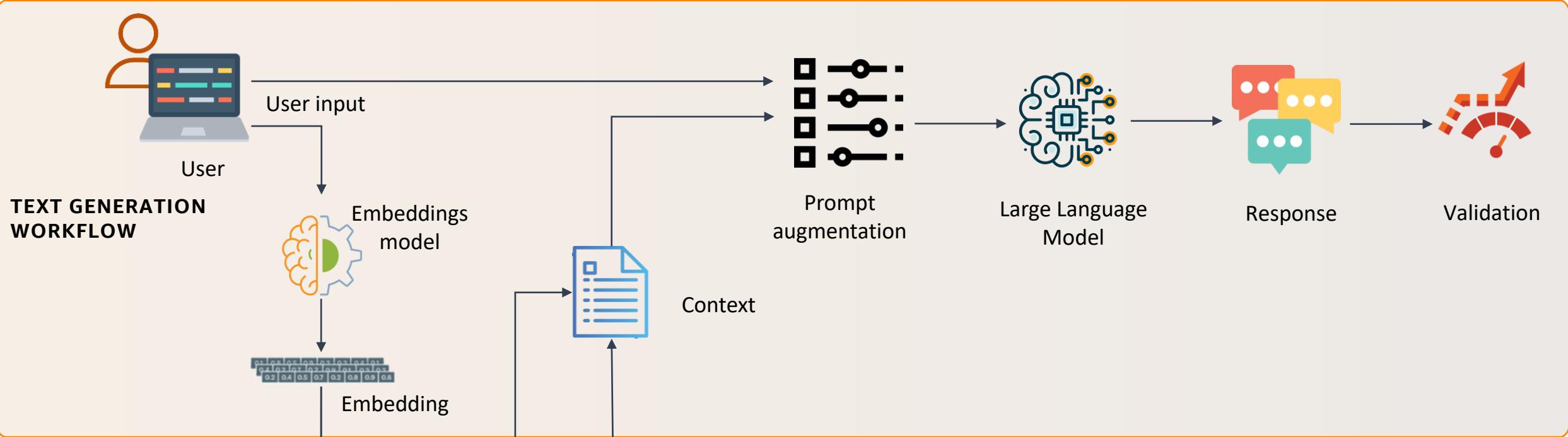
\* Both authors contributed equally to this research.

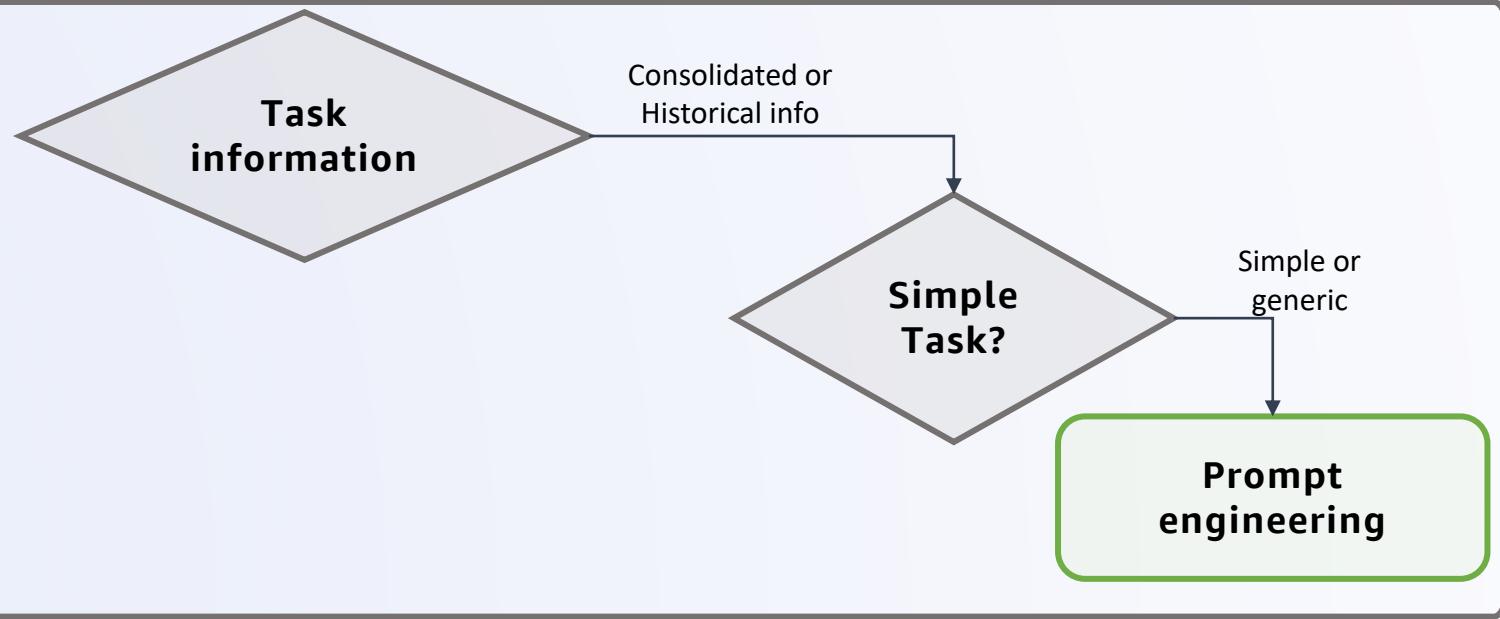
## ABSTRACT

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%.

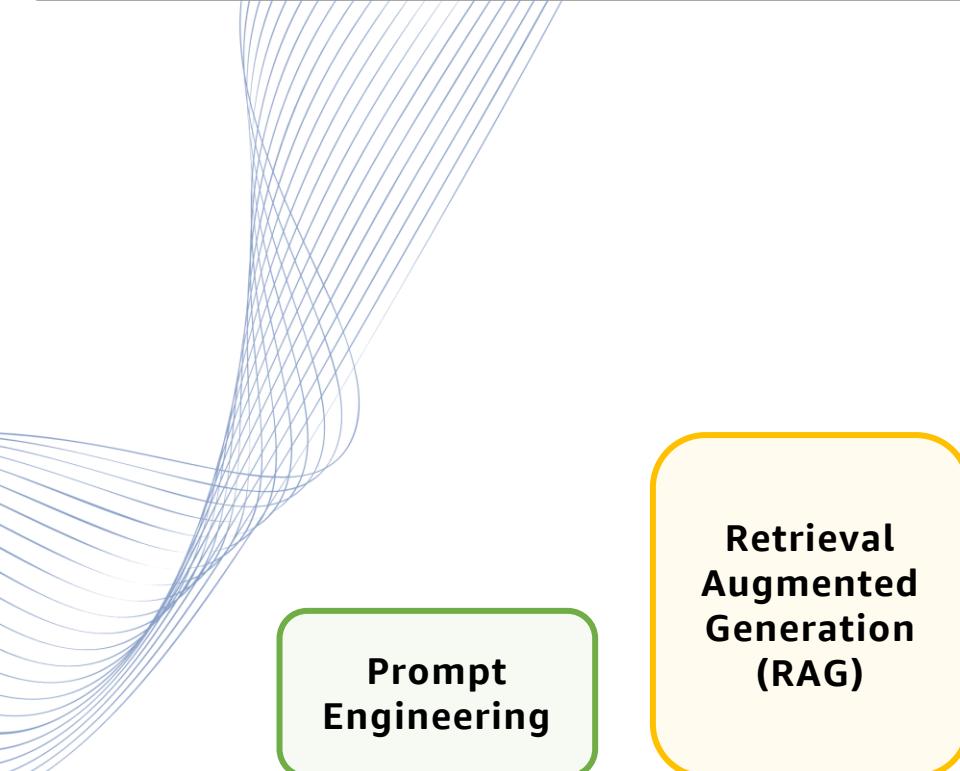
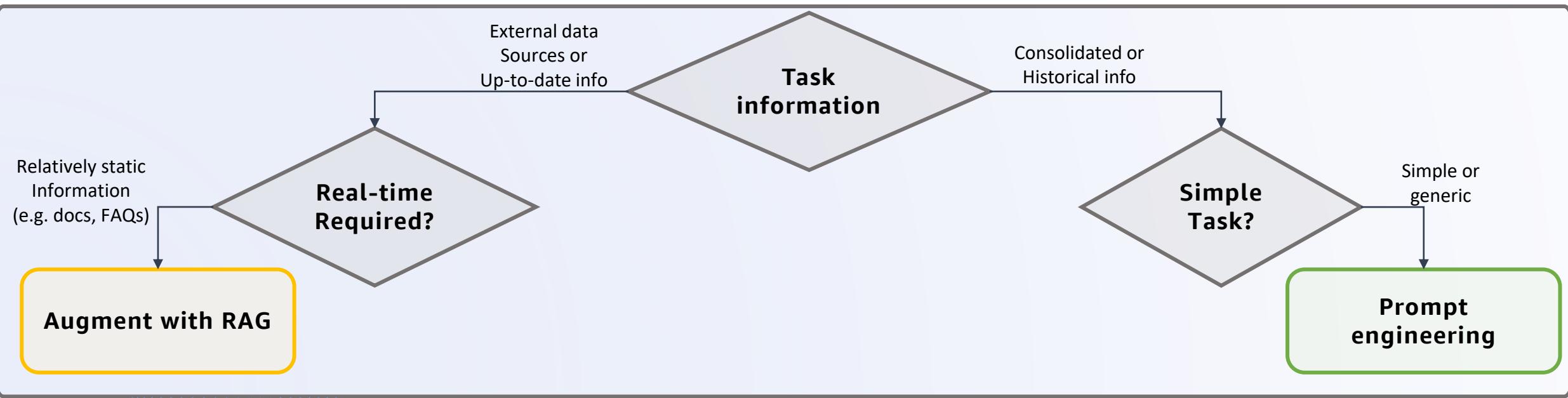
<https://arxiv.org/abs/2401.07883>

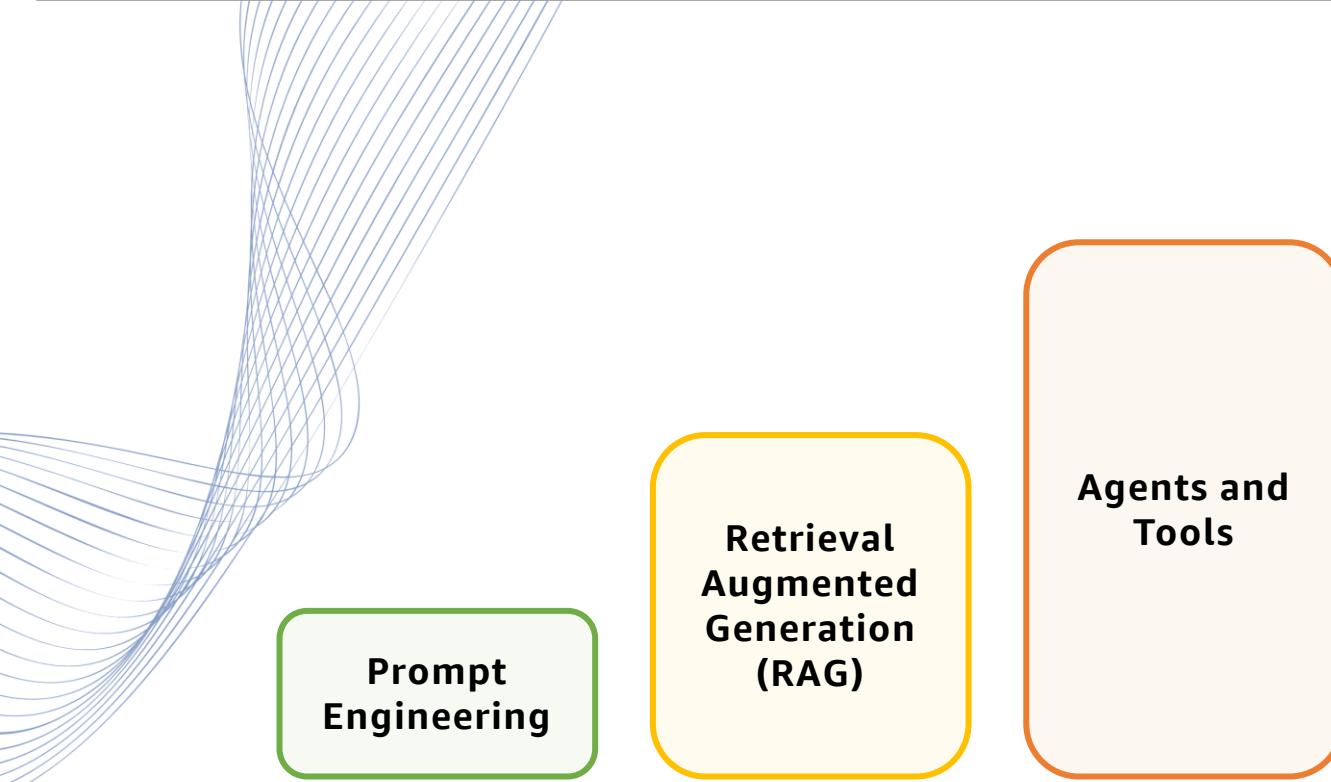
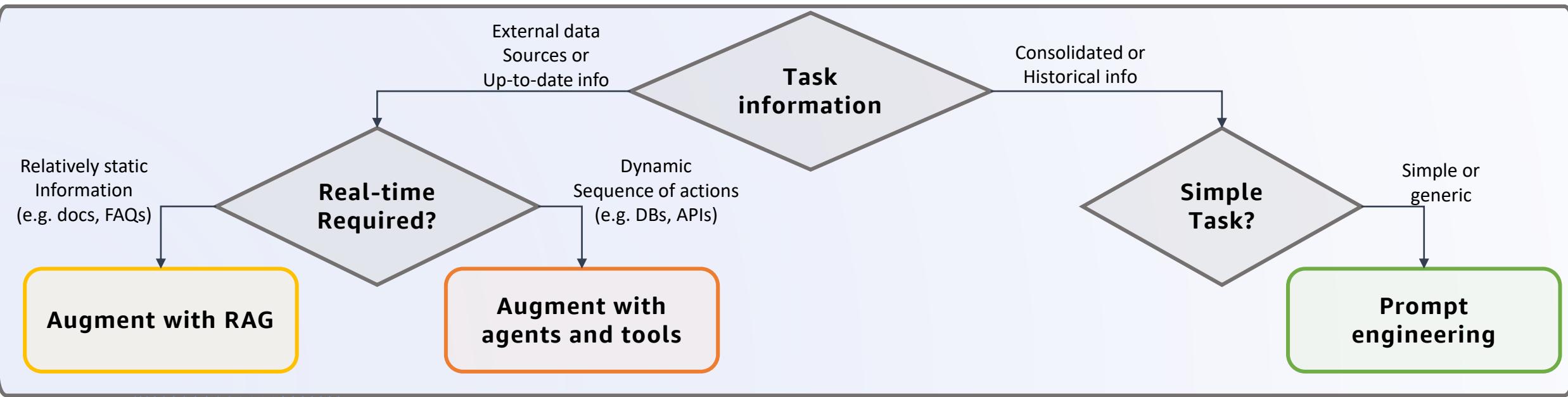


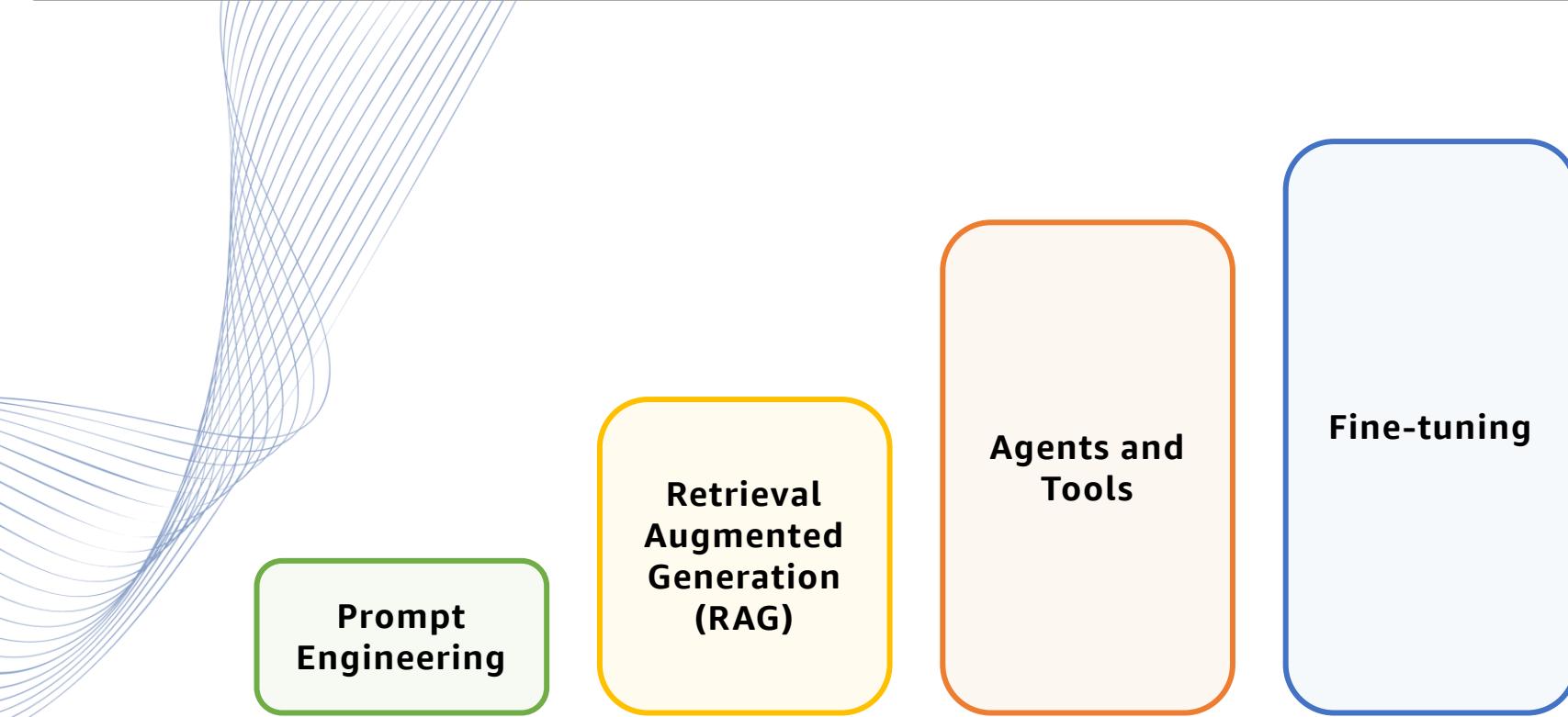
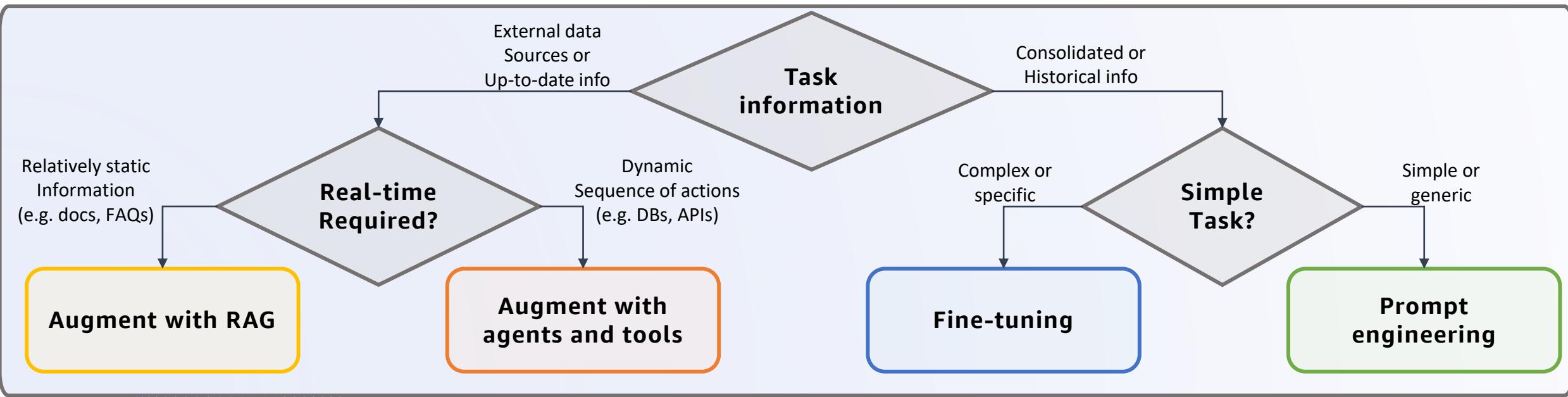


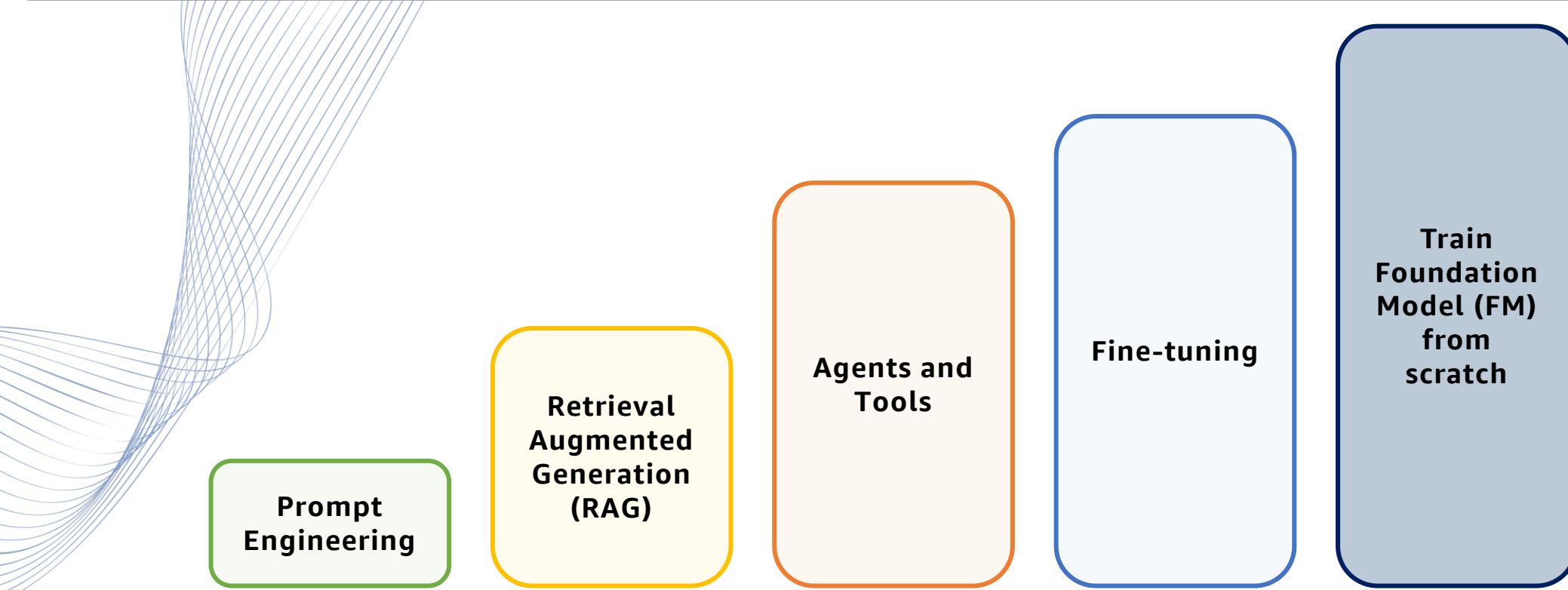
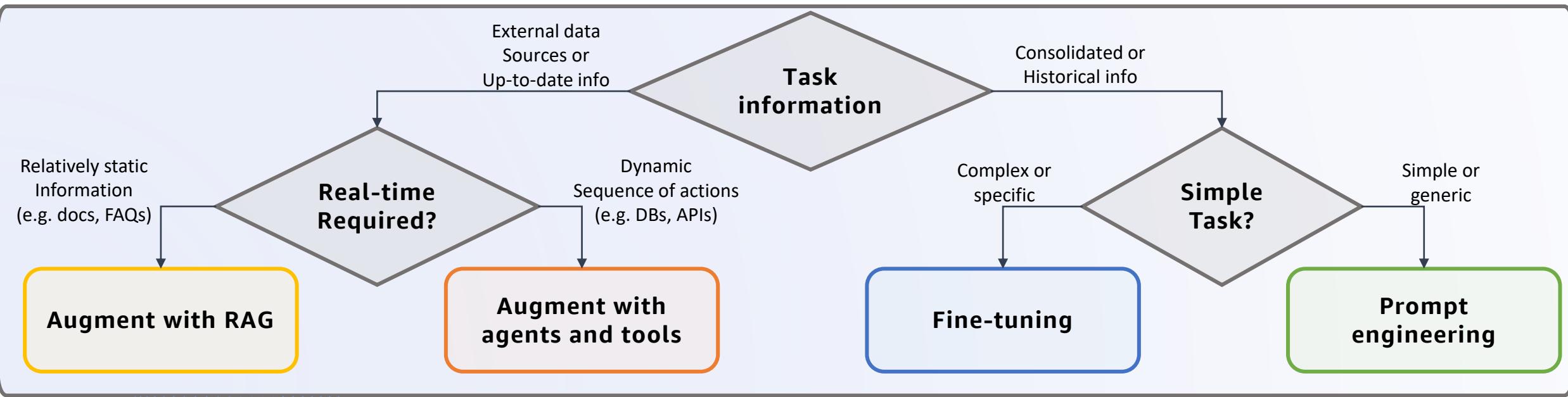


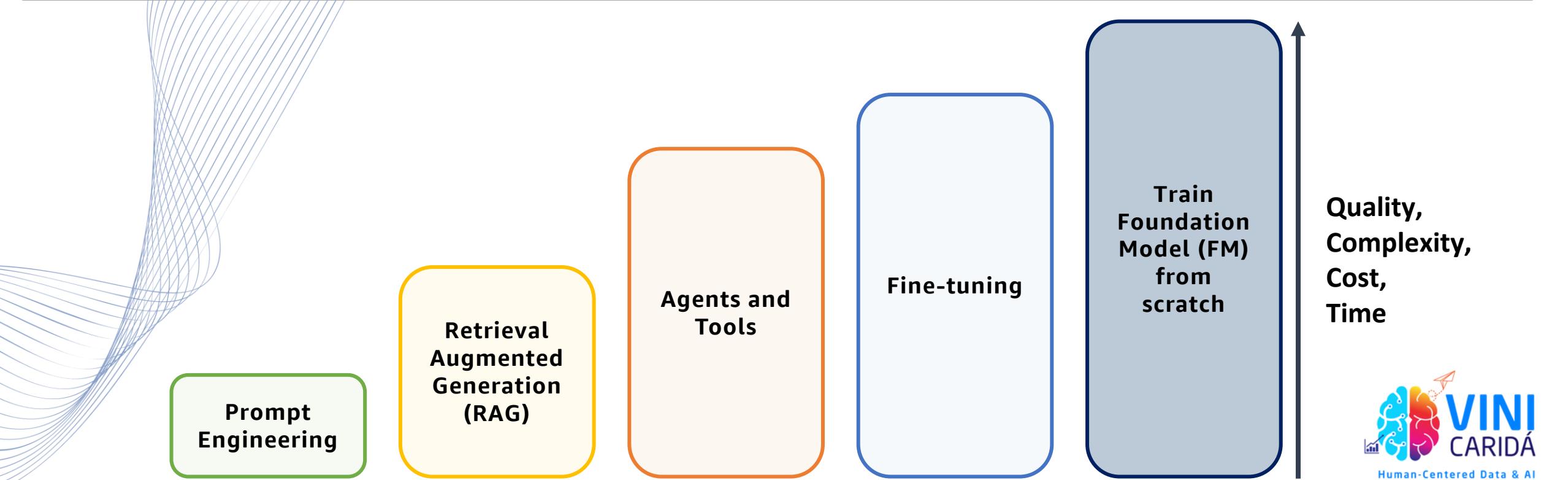
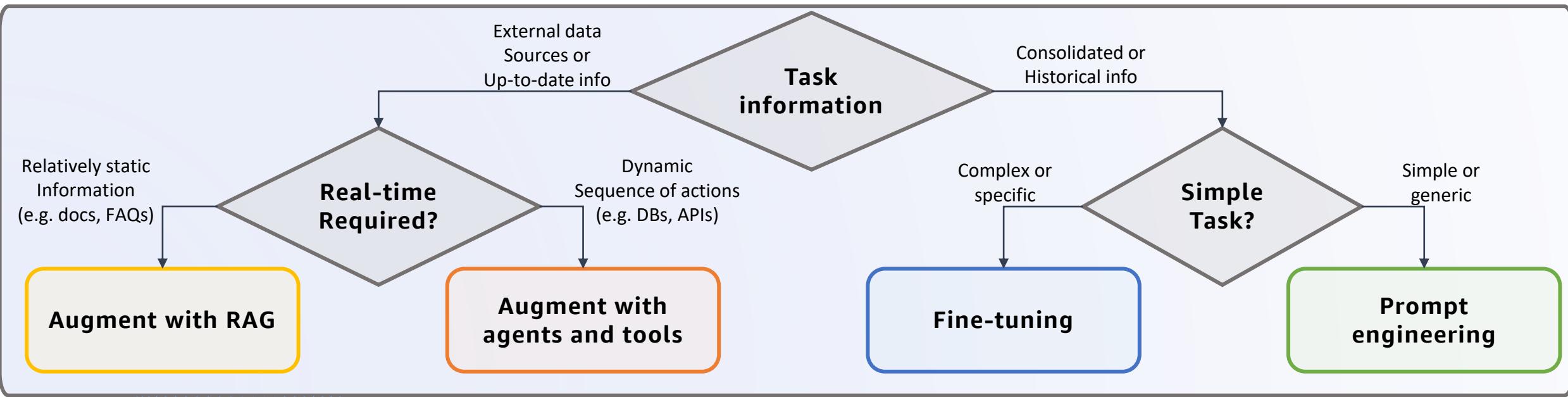
Prompt  
Engineering

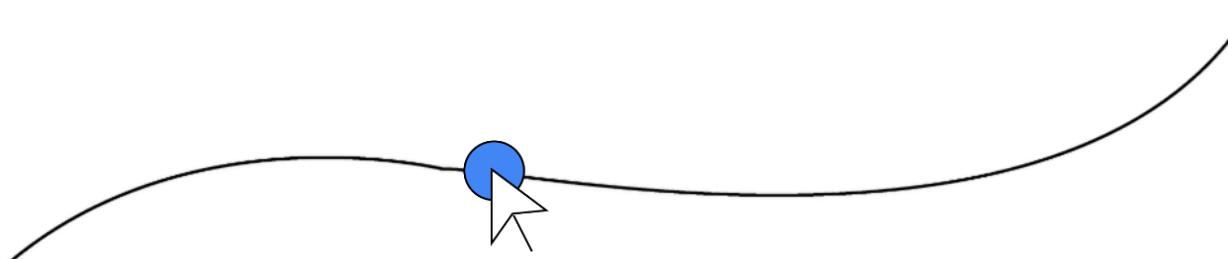












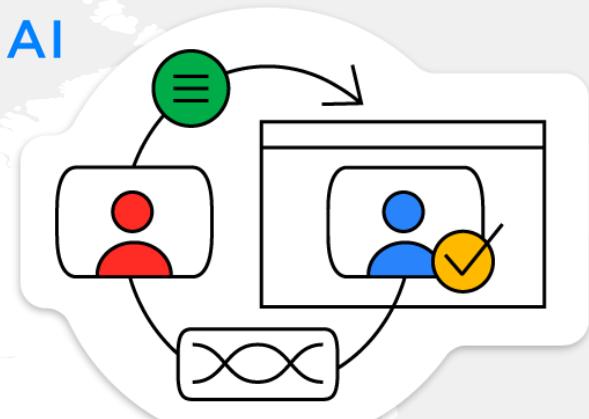
# *Bora ver na prática!*





**VINICARIDÁ**  
Human-Centered Data & AI

# Thank you!



@vinicius caridá



@vfcarida



vfcarida@gmail.com



<https://linktr.ee/vfcarida>