

# Preparando-se para o Futuro

AI Leadership Navigator  
**HORIZONTES DA AI NOS NEGÓCIOS**



# Alan Braz

Xillenial 1980

Campinas

Sofredor do Guarani

Unicamp - CC MSc

2 filhos - Luiza e Mateus

Esportes NBA/NFL

Fx preta TKD

Empolgado!



[alanbraz.com.br](http://alanbraz.com.br) / [pullrecast.dev](http://pullrecast.dev)



# Internet



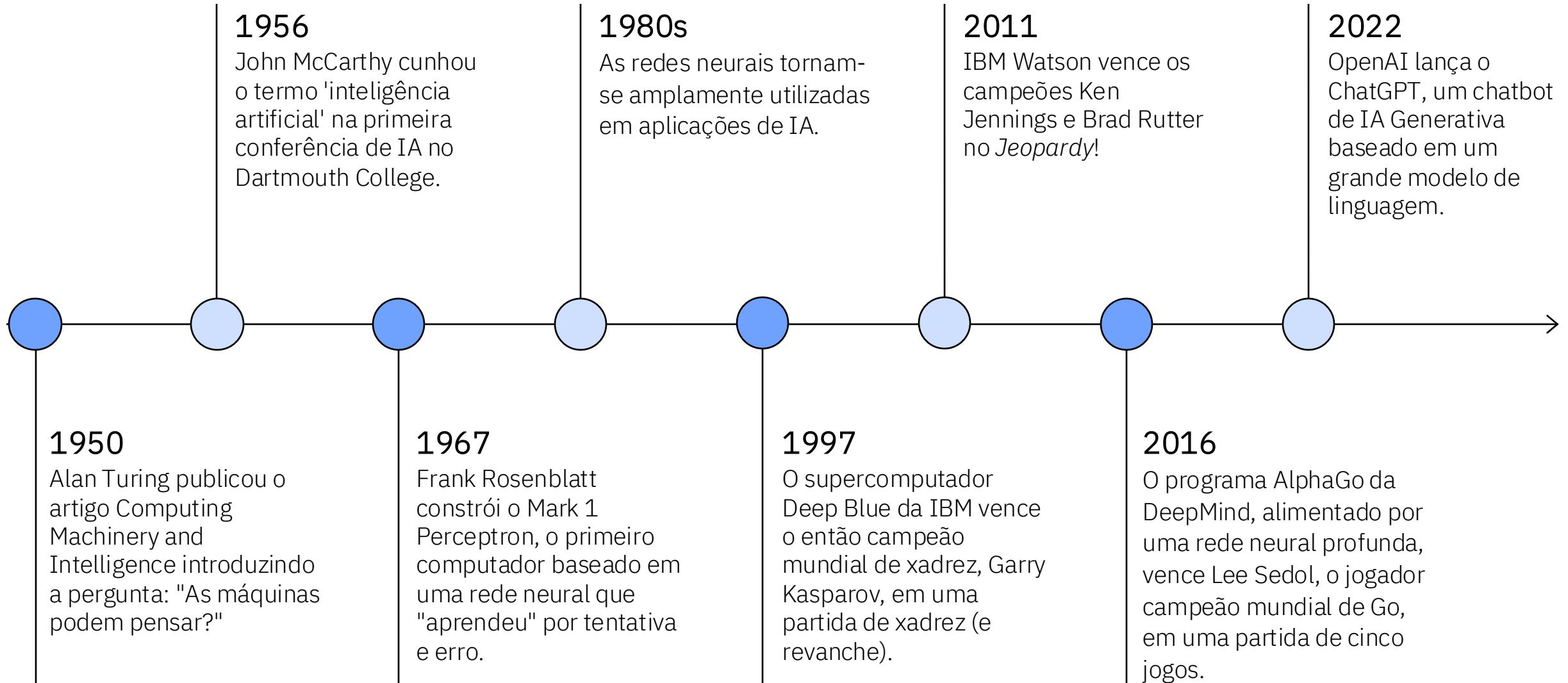
# IA





[alanbraz.com.br](http://alanbraz.com.br) / [pullrecast.dev](http://pullrecast.dev)

# Marcos da IA



# Deep Blue 1997



# Watson 2011



# Ei! 2013/2014



# Chef Watson 2017



# 2019

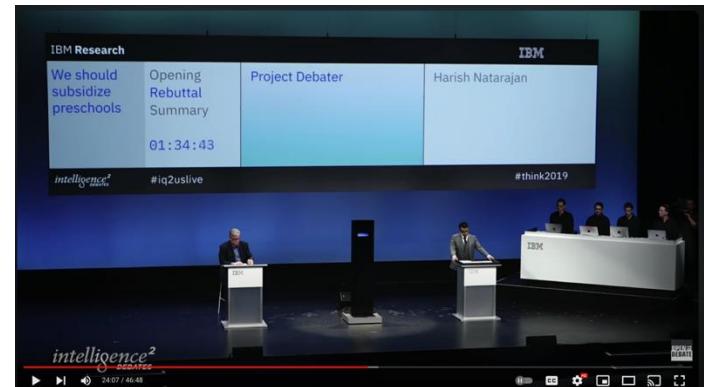
## O Boticário lança 1ºs perfumes feitos com ajuda de inteligência artificial

Em investida inédita no mundo, fragrâncias foram desenvolvidas em parceria com a IBM e chegam ao mercado no dia 27. Conheça em primeira mão a novidade

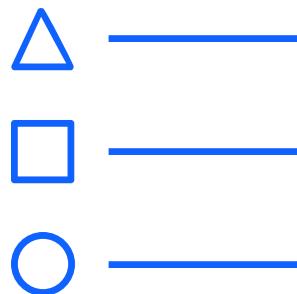


Marco na perfumaria: fragrâncias criadas com ajuda de "robô" chegam dia 27. (Grupo Boticário/Divulgação)

# Project Debater



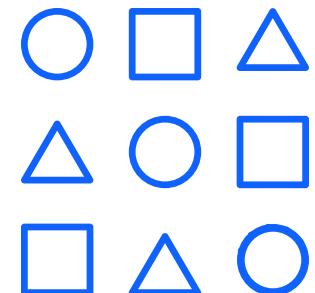
## Sistemas baseados em regras



*programado com uma série de instruções e regras lógicas*

*Por exemplo:  
"um triângulo tem três lados"*

## Sistemas de aprendizado de máquina



*treinado com uma série de exemplos*

*Por exemplo: "Aqui estão fotos de muitos triângulos"*

# Inteligência Artificial (IA)

*Inteligência humana imitada por máquinas*



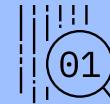
## Machine Learning (ML)

*Sistemas que aprendem com dados históricos*



## Deep Learning (DL)

*Técnica de ML que imita a função cerebral humana*



## Foundation Model

*Sistemas de IA Generativa*



Modelo de IA criado usando um tipo específico de arquitetura de rede neural, chamado Transformer, que é projetado para gerar sequências de elementos de dados relacionados (por exemplo, como uma frase).



1950's



1980's



2010's



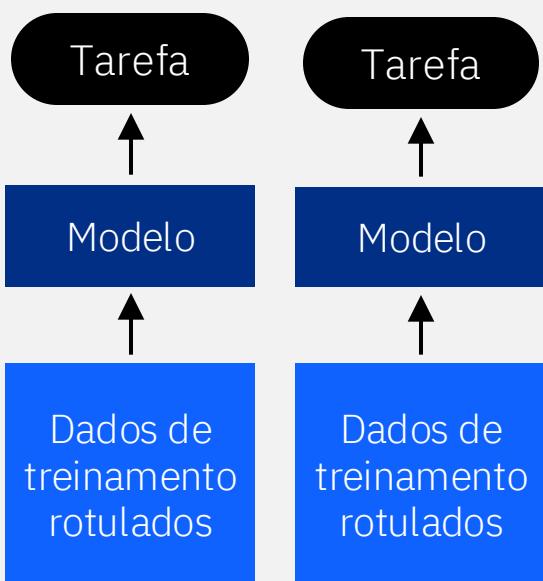
2020's



## O que vem a seguir com os modelos de fundação:

IA convencional

1000s – 100000s Pontos de dados rotulados / Tarefa



Zero-shot prompting

No labeled data needed

Tarefa



Few-shot prompting

1-10 labeled data points / task

Tarefa



Prompt-tuning

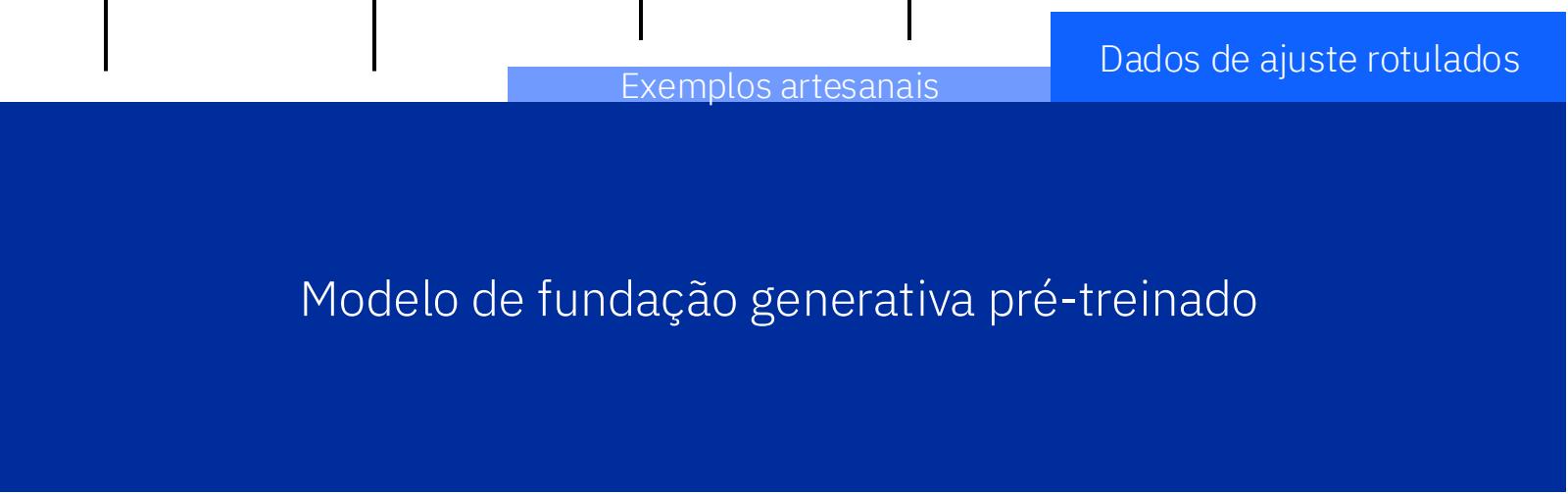
100s – 1000s labeled data points / task

Tarefa

Tarefa

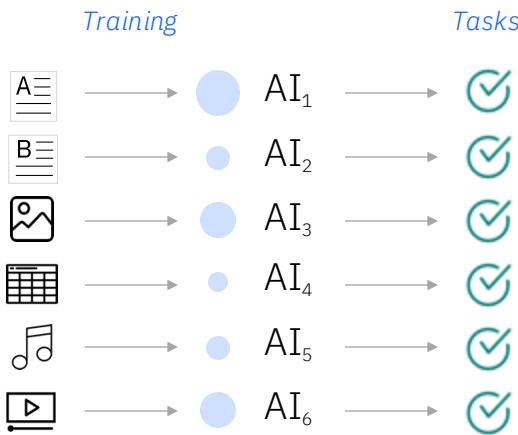


Modelo de fundação generativa pré-treinado



# Foundation models establish a new paradigm for AI capabilities

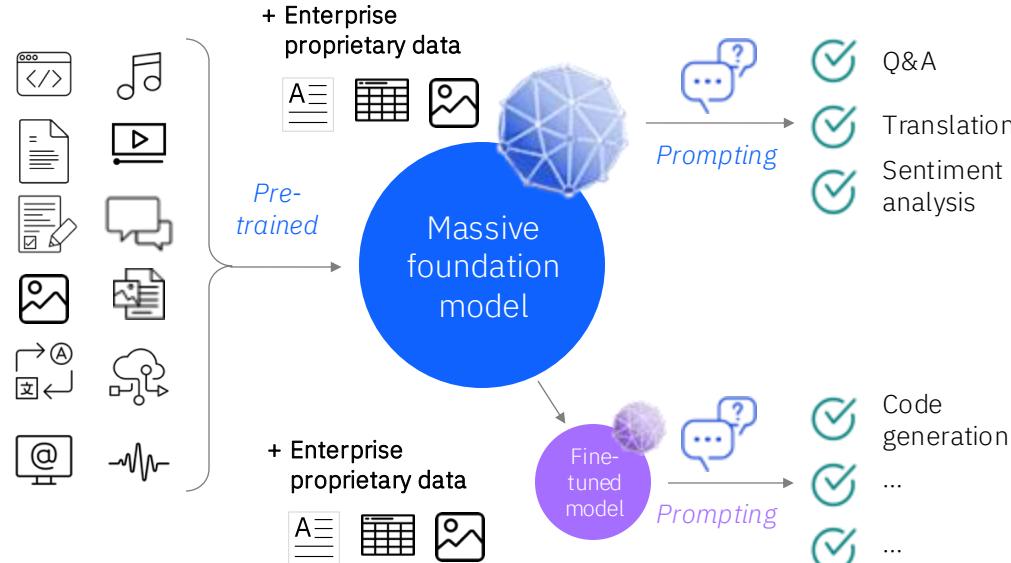
## Traditional AI models



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

## Foundation models

External data



- Massive multi-tasking model
- Adaptable with minimized training
- Pre-trained unsupervised learning

## Enhanced capabilities

- Summarization
- Conversational knowledge
- Content creation
- Code co-creation

## Key advantages

- Lower upfront costs through less labeling
- Faster deployment through fine tuning
- Equal or better accuracy for multiple use cases
- Incremental revenue through better performance

up to **70% reduction** in certain NLP tasks

# 82%

Líderes empresariais dizem que a complexidade da TI está impedindo o sucesso

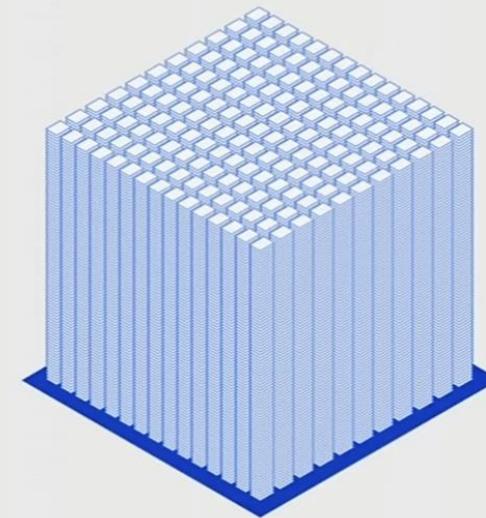
# 55%

Líderes empresariais sem informações importantes sobre seus gastos com tecnologia

# 10%

Empresas em modo de produção com IA generativa

Aproximadamente todos os dados disponíveis publicamente estão representados em modelos fundacionais



Menos de 1% de todos os dados empresariais representados nos modelos fundacionais



## Tendências de mercado ↴

### Criação de Valor

- As disruptões lideradas pela tecnologia estão aceleradas, impulsionadas pela IA generativa.
  - A experiência é mais importante do que pensamos.
  - Novos produtos, serviços e modelos de negócios.
- 

### Produtividade

- As promessas de sustentabilidade estão apenas começando a mostrar progresso.
  - Novos modelos operacionais mais eficientes.
  - A reglobalização, a requalificação e as novas parcerias estão transformando as cadeias de valor.
- 

### Talento

- Novos perfis de liderança de impacto.
- Cultura mais colaborativa, ágil e digital.
- A escassez de competências e mudanças demográficas demandam um novo contrato social.

# IA para negócios

A **precisão** por si só não é mais suficiente. Uma IA **confiável** é crucial.



## Justa

Treinada em dados imparciais.



## Transparente

Aberta à inspeção e avaliação.



## Explicável

Para que os usuários entendam facilmente os resultados e decisões indicados pela IA.



## Robusta

Elimina os riscos de segurança e permite a confiança em resultados de sistemas de IA.



## Privada

Alimentada por dados de alta integridade compatíveis com os negócios e defende a privacidade e necessidades de segurança.

EFICIENTE

Reinventando a forma como o trabalho é feito

+ AI → AI +

*Reinventando a forma como o trabalho  
é feito em todos os domínios de  
negócios e setores*

# A escada da IA moderna

+AI



Colete, organize e aumente os dados



Adicione IA aos seus aplicativos



Automatize seus fluxos de trabalho



Substitua seus fluxos de trabalho



A IA faz o trabalho

AI+

“AI First” ↗

## Computação Quântica

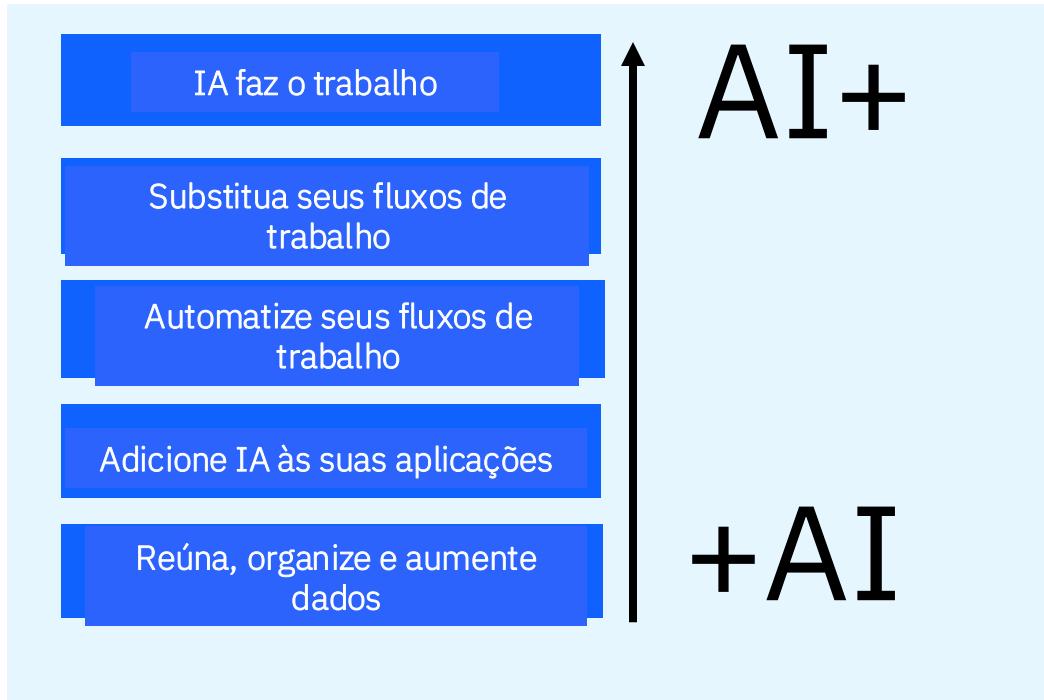
Melhora significativamente a capacidade de executar cargas de trabalho de IA

## Hiperautomação

87% dos executivos dizem que IA ajuda a executar mais iniciativas de automação de alto impacto

## Robótic

Aspera-se que a integração de tecnologias emergentes impulsione a robótica para novas fronteiras. O mercado de robótica de IA deve atingir US\$ 19,01 bilhões em 2024



## Cibersegurança

84% planejam priorizar soluções de cibersegurança de IA Gen em relação às soluções convencionais de segurança

## Infraestrutura /Nuvem Híbrida

A IA acelera jornadas de nuvem híbrida com 53% de melhoria no desempenho de TI ou rede com IA

## Realidade Aumentada

Gera economia de até 40% em produtividade

## Data & Analytics

+50% CEOs enxergam Analytics como uma tecnologia-chave para melhorar seus negócios

## IOT

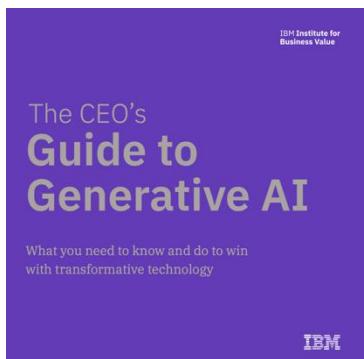
A IA aprimora os recursos dos sistemas IoT processando e analisando dados gerados em tempo real

## IBM POV:

### Quatro princípios fundamentais para adaptar a IA generativa para empresas

#### Aberta

- Baseado nas melhores tecnologias de IA e nuvem disponíveis
- Dar acesso à inovação da comunidade aberta e a múltiplos modelos



#### Focada

- Projetado para casos de uso de negócios direcionados, que desbloqueiam um novo valor
- Incluindo modelos selecionados que podem ser ajustados a dados proprietários e diretrizes da empresa

#### Confiável

- Construído com IA e governança de dados, transparência e ética que suportam as crescentes demandas de conformidade regulatória
- Fornecer orientação sobre modelos apropriados para alavancar para criar valor comercial real com confiança

#### Capacitante

- Em uma plataforma para trazer seus próprios dados e modelos de IA que você ajusta, treina, implanta e controla
- Execução em qualquer lugar, projetado para escala e adoção generalizada para realmente criar valor empresarial

<https://link.pullrecast.dev/ceo-guide-genai>

# Employee Experience ↗



Um único canal:  
experiência  
envolvente

2 camadas: Digital &  
Humano



4,7k páginas  
de conteúdo de RH



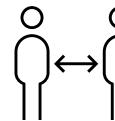
243k  
Usuários  
únicos



97%  
Adoção de gerentes



50+  
Idiomas suportados



1.9m  
Interações



94%  
Adoção de  
executivos

Profissionais de RH atuando em atividades estratégicas

---

75%

Melhora em  
produtividade

61%

Redução de  
abertura de  
chamados

75%

Feedbacks  
positivos  
(CSAT)

89%

de interações  
resolvidas  
diretamente via  
AskHR

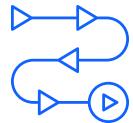
# O melhor da IA

Quando usar

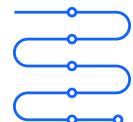
## Recursos de IA **Tradicional**



Preditivo/Prescritivo



IA conversacional dirigida



IA de Visão Dirigida

Quando usar

## Recursos de IA **Generativa**



Sumarização



Pesquisa conversacional

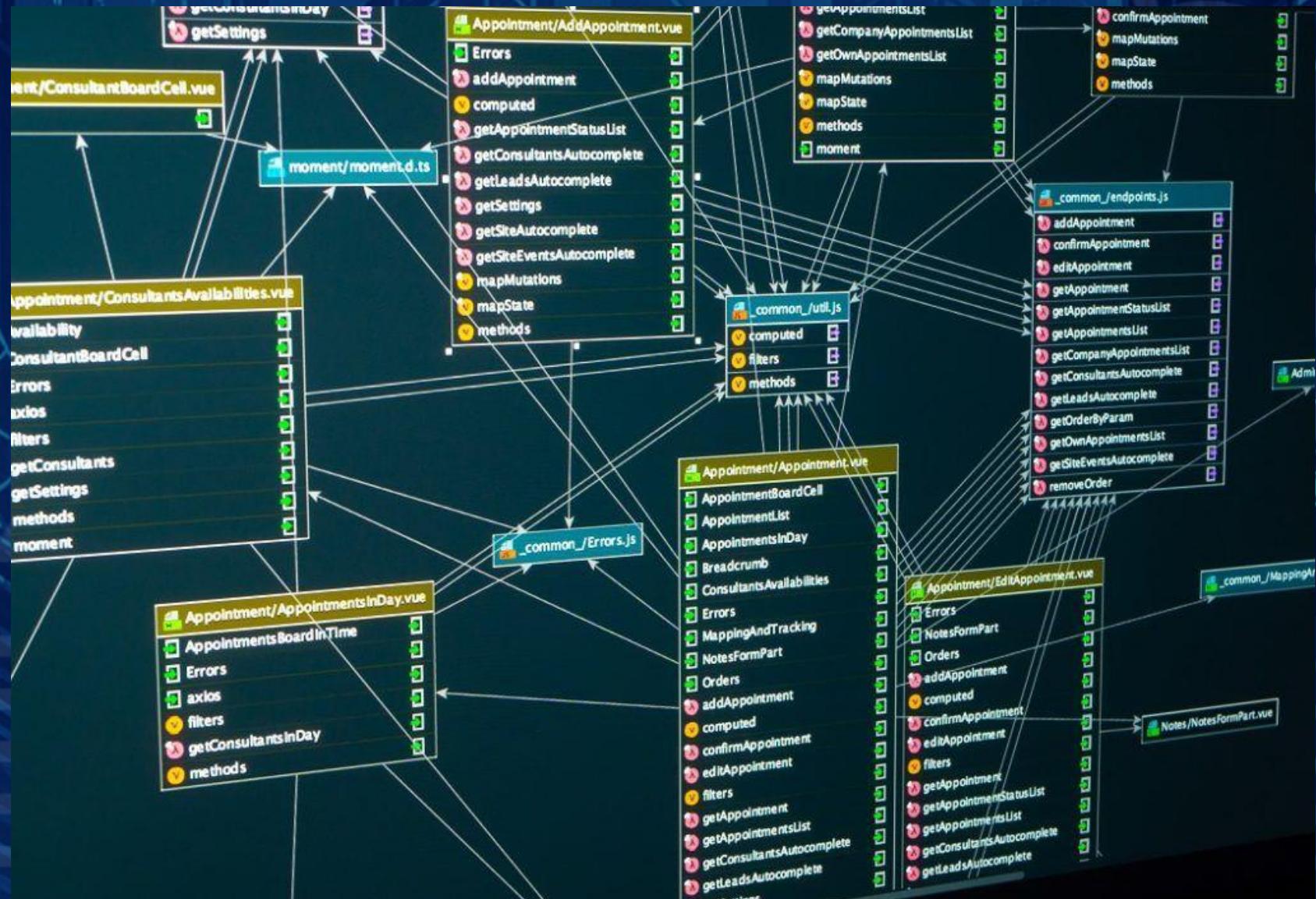


Criação de conteúdo



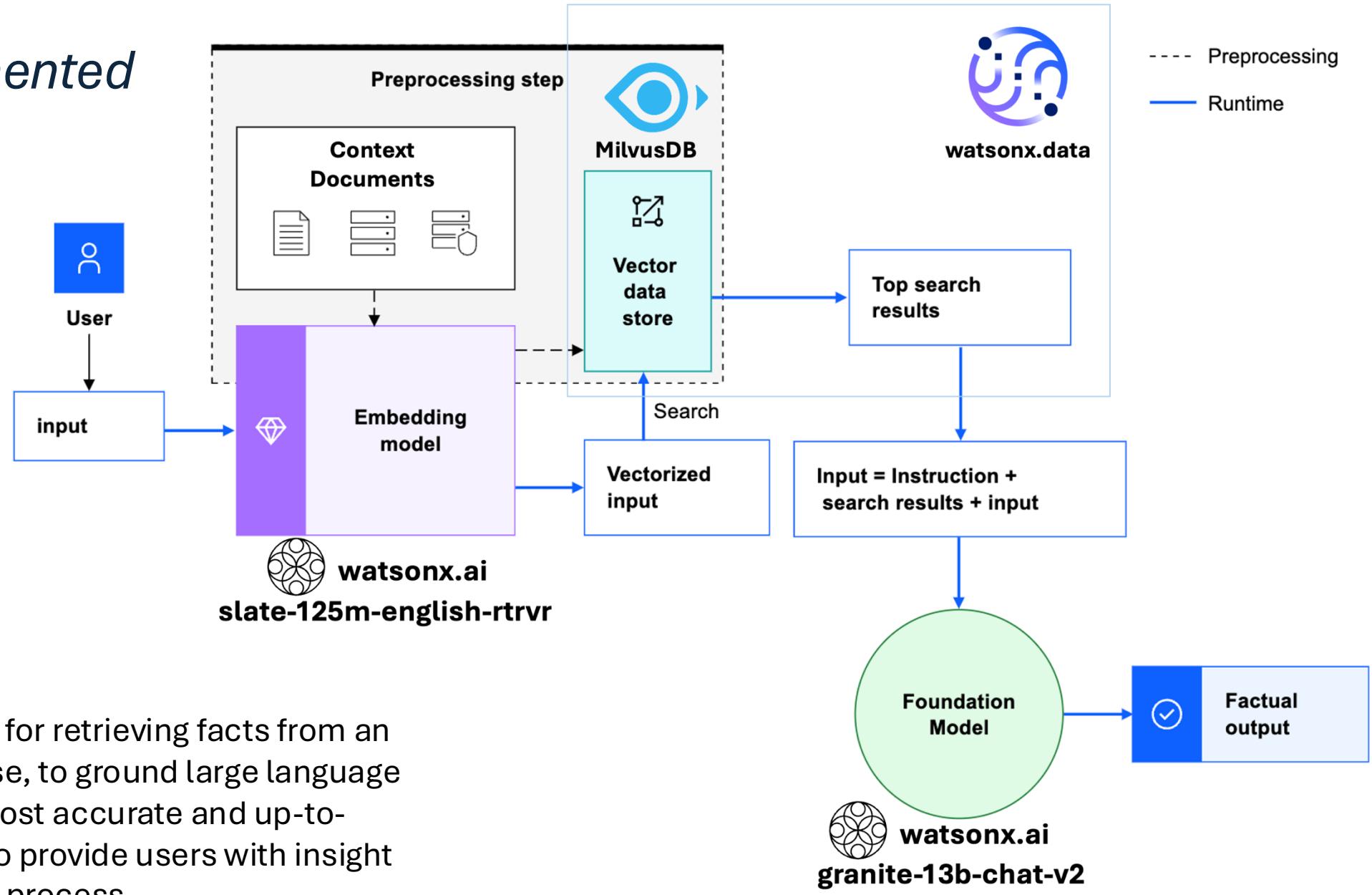
Criação de código

LLM



# RAG

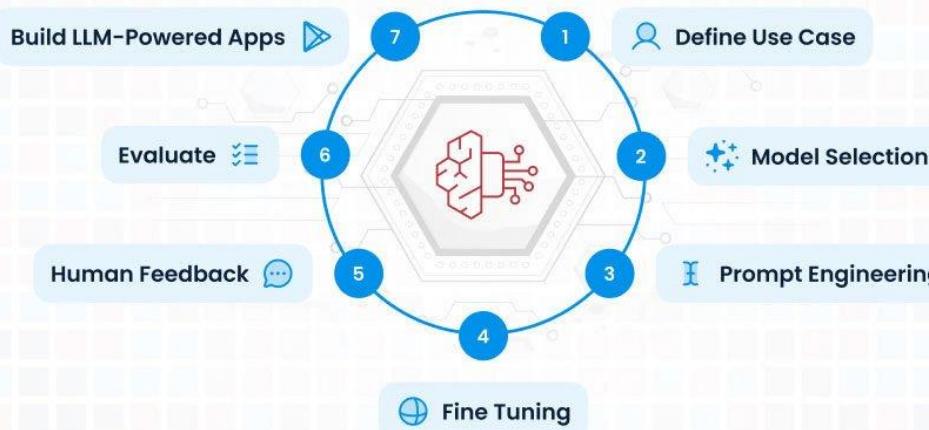
## Retrieval Augmented Generation



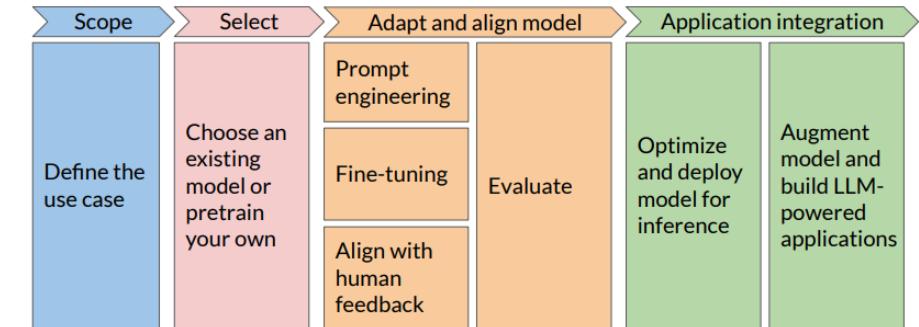
RAG is an AI framework for retrieving facts from an external knowledge base, to ground large language models (LLMs) in the most accurate and up-to-date information, and to provide users with insight into the LLM generating process.

# Inovação

## Gen AI Project Lifecycle



## Generative AI project lifecycle



# IBM Research

IBM Research is home to a community of scientists, engineers, and designers. We are drawn together by a passion to imagine the future of computing, and to build it.

## Semiconductors

Artificial intelligence

Africa

Ireland

Quantum computing

Albany  
Almaden

Israel  
Tokyo

Hybrid cloud

Brazil

United Kingdom  
Cambridge

Security

India

Yorktown Heights  
Zurich



6 Nobel Laureates



10 Medals of Technology



5 National Medals of Science



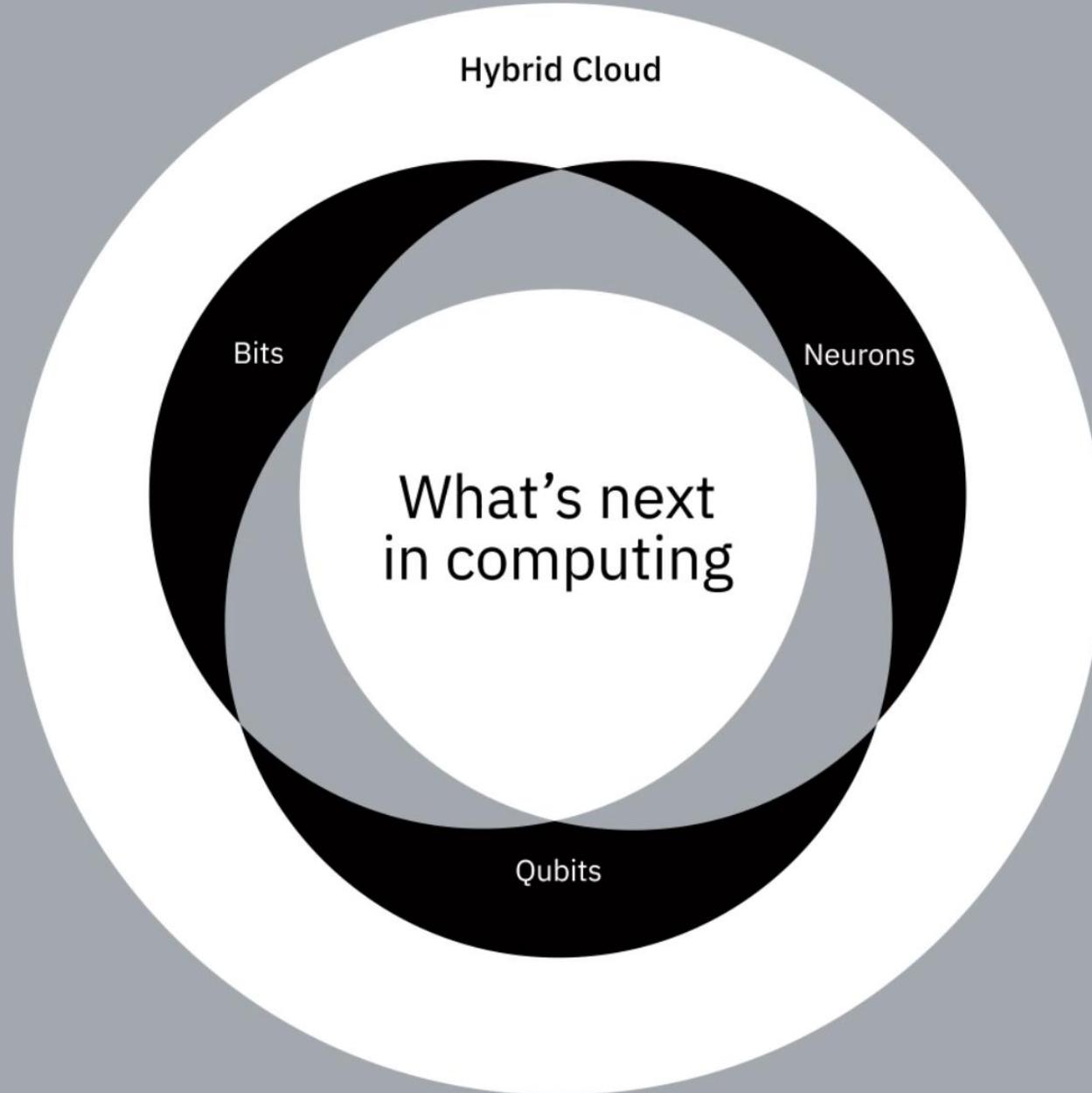
6 Turing Awards











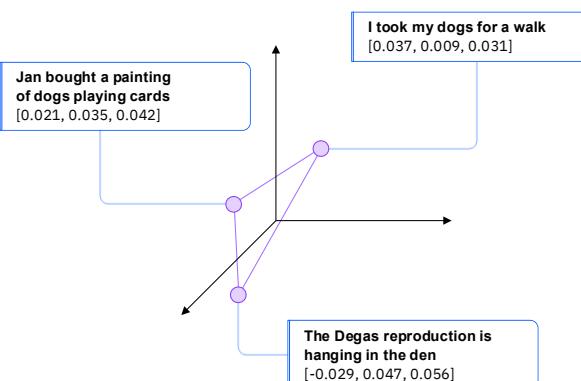
# AI Platform Engineering

- Time global de Engenharia de Software em Research
- Faz a ponte entre pesquisa e produto (IBM Software, Red Hat, Infra.)
- Temas do grupo no Brasil [BR](#)

Guardrails



Embedding



Otimizações na inferência (OS)

vLLM  
TGI

Compilador e performance em AIU



<https://research.ibm.com/blog/ibm-artificial-intelligence-unit-aiu>

# 2017

## AI Model Catalog

≡ Models list Stats User guide About

Search models by name, description, domain type, tag name, defined, training, trained, valid or invalid.

129 public models found.

**Filters:**

- > Lifecycle Stages
- > Readiness
- ▽ Domain
  - Computer Vision
  - Graphs
  - Medical (129)
  - Miscellaneous
  - Natural Language Processing
  - Playing Games
  - Speech
  - Unknown
- ▽ Frameworks
  - Caffe (1)
  - Caffe2
  - Keras (43)
  - Lasagne (2)
  - MXNet (1)
  - NLTK (2)
  - PyTorch (30)
  - TensorFlow (61)
  - Theano (8)

Name	Domain	Framework(s)	Lifecycle stages	Readiness	Created at ↓
# ★ 190	Medical	PyTorch	Definition (70%)	Trainability (35%)	16 Aug 2019
CSCI5922_Project ★ 0	Medical	TensorFlow, Keras	Definition (70%)	Trainability (35%)	16 Aug 2019
kaggle-tgs-salt-competition ★ 1	Medical	TensorFlow, Keras	Definition (70%)	Trainability (35%)	16 Aug 2019
Data Science Bowl 2018 ★ 0	Medical	TensorFlow, Keras	Definition (70%)	Trainability (35%)	16 Aug 2019
HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images ★ 17	Medical	TensorFlow	Definition (70%)	Trainability (50%)	16 Aug 2019
HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images ★ 17	Medical	TensorFlow	Definition (70%)	Trainability (50%)	16 Aug 2019
Project Overview ★ 13	Medical	TensorFlow, Keras	Definition (70%)	Trainability (35%)	16 Aug 2019
PSCN is a python3 implementation of the paper "Learning Convolutional Neural Networks for Graphs" by Mathias Niepert, Mohamed Ahmed and Konstantin Kutzkov ( <a href="https://arxiv.org/abs/1605.05273">https://arxiv.org/abs/1605.05273</a> ) ★ 19	Medical	TensorFlow	Definition (70%)	Trainability (50%)	16 Aug 2019
RedeemTheBoar ★ 1	Medical	TensorFlow, Keras	Definition (70%)	Trainability (35%)	16 Aug 2019
cv_course_project ★ 1	Medical	PyTorch, scikit-learn	Definition (70%)	Trainability (50%)	16 Aug 2019
Unet_keras ★ 0	Medical	TensorFlow	Definition (70%)	Trainability (35%)	16 Aug 2019
Complementary_Segmentation_Network (Outperforms u-nets everytime :) for binary segmentation ) ★ 15	Medical	TensorFlow, Keras	Definition (70%)	Trainability (50%)	16 Aug 2019
Complementary_Segmentation_Network (Outperforms u-nets everytime :) for binary segmentation ) ★ 15	Medical	TensorFlow, Keras	Definition (70%)	Trainability (50%)	16 Aug 2019
Complementary_Segmentation_Network (Outperforms u-nets everytime :) for binary segmentation ) ★ 15	Medical	TensorFlow, Keras	Definition (70%)	Trainability (50%)	16 Aug 2019
Note: This README is an early work-in-progress. ★ 1	Medical	TensorFlow	Definition (70%)	Trainability (35%)	16 Aug 2019



# Hugging Face

 Hugging Face  Models Datasets Spaces Community Docs Pricing

[+ New](#)

alanbraz

- Profile
- Inbox (0)
- Settings
- Billing
- Get **Pro**

Organizations

- IBM
- Create New

Resources

- Hub guide
- Transformers doc
- Forum
- Tasks
- Learn

Light theme

**Following** 2

All **Models** Datasets Spaces Papers Collections Community Posts

Upvotes Likes Articles

ibm-research models updated by SagiPolaczek 5 months ago

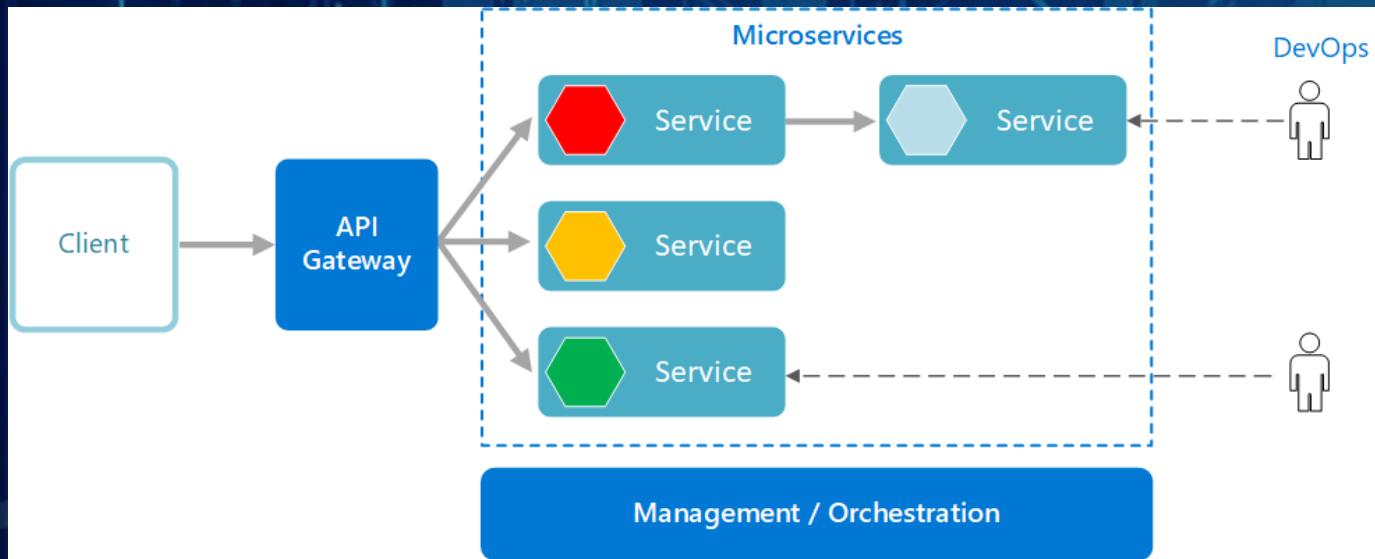
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.protein\_soluti... Updated Apr 5 • ↓ 19 • ❤ 5
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.dti\_bindingdb\_... Updated Dec 19, 2024 • ↓ 24 • ❤ 2
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.tcr\_epitope\_b... Updated Dec 19, 2024 • ↓ 9 • ❤ 2
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.moleculenet\_cl... Updated Dec 19, 2024 • ↓ 6
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.moleculenet\_cl... Updated Dec 19, 2024 • ↓ 8
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.moleculenet\_bb... Updated Dec 19, 2024 • ↓ 15 • ❤ 1
- ibm-research/biomed.omics.bl.sm.ma-ted-458m.dti\_bindingdb\_... Updated Dec 19, 2024 • ↓ 6
- ibm-research/biomed.omics.bl.sm.ma-ted-458m

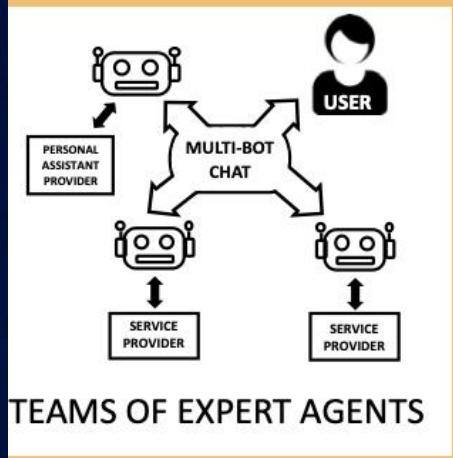
**Trending** last 7 days

All Models Datasets Spaces

- mistralai/Devstral-Small-2505 Text2Text Generation Updated ... ↓ 45.9k ❤ 479
- google/gemma-3n-E4B-it-litert-preview Image-Text-to-Text Updated 4 days ago ↓ 436
- ByteDance-Seed/BAGEL-7B-MoT Any-to-Any Updated 2 days ago ↓ 998 ❤ 404
- DeepSite** Generate any application with DeepSeek
- LTX Video Fast ultra-fast video model, LTX 0.9.7 13B distilled
- google/medgemma-4b-it Image-Text-to-Text Updated 3 d... ↓ 8.49k ❤ 172
- FLUX Pro Unlimited** Use the FLUX-Pro model as much as you want.
- disco-eth/EuroSpeech Viewer Updated 6 day... ↓ 8.42M ↓ 27.4k ❤ 52

# Agentes de IA





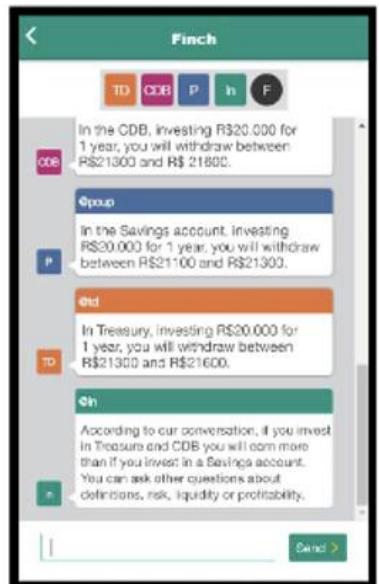
**cdbGuru**  
reactive bot



**poupancaGuru**  
reactive bot



**tesouroGuru**  
reactive bot



**M** User

**investmentGuru**  
proactive bot



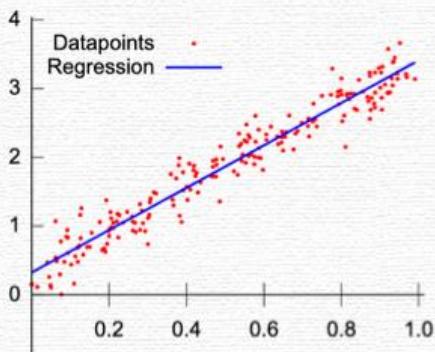
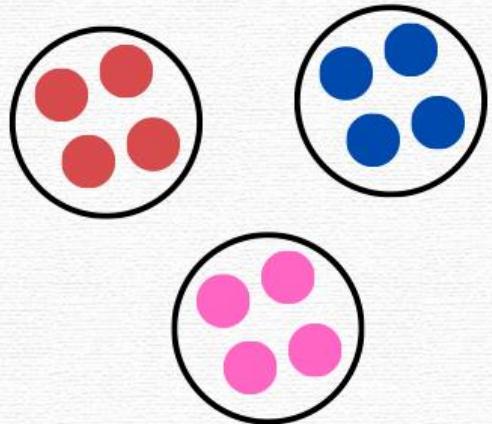
from Q&A dialog to multiparty chat

Self-service finance advising  
using multiple chatbots

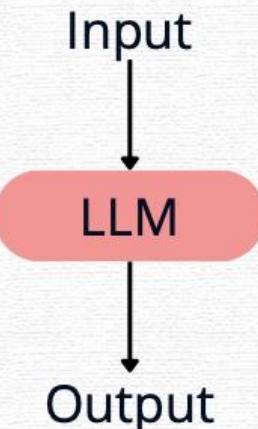
Should we design multi or single bots?  
[DeepDial'18]

# Evolução da Inteligência Artificial

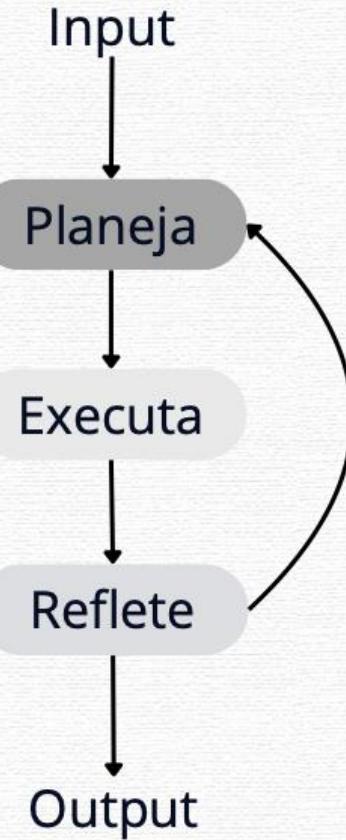
➤ Classificação/Predição



➤ Assistants



➤ AI Agent

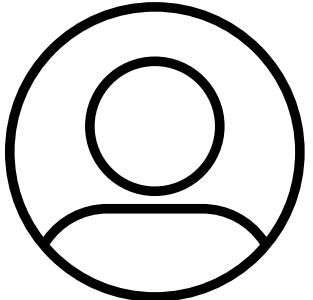


# Introdução

## O que são AI Agents ?

Um agente de IA é um sistema autônomo que pode usar **ferramentas** e **colaborar** com outros agentes para **planejar** e **executar** tarefas. Após agir, o agente reflete sobre os resultados de suas ações, **aprendendo** de forma iterativa e refinando sua abordagem para alinhar-se melhor com seus objetivos definidos.





Role Playing

Focus

Tools

Cooperation

Guardrails

Memory

# Quick demo - Full visibility of the interim steps

**User:** What's the current weather in Las Vegas?

**Agent:** The current weather in Las Vegas is 72.1°F with an apparent temperature of 64°F.

**User:** What's the current weather in Las Vegas?

**Agent: (thought)** I can use the OpenMeteo tool to get the current weather in Las Vegas.

Agent: (tool\_name) OpenMeteo

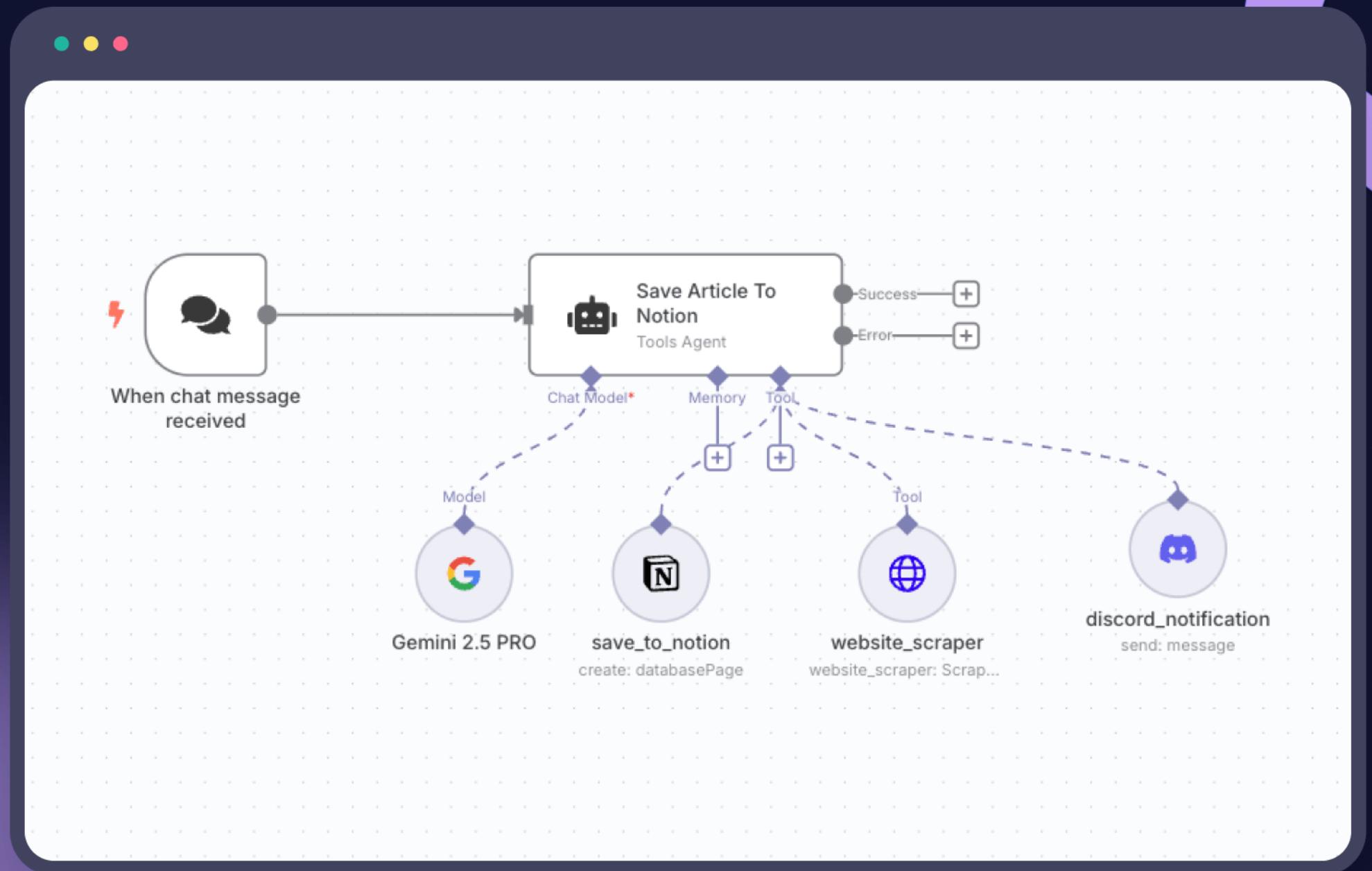
Agent: (tool\_caption) Getting current weather for Las Vegas

```
Agent: (tool_input) {"location": {"name": "Las Vegas", "language": "English"}, "start_date": "2024-09-13", "end_date": "2024-09-13", "temperature_unit": "celsius", "units": {"time": "iso8601", "apparent_temperature_max": "°C", "apparent_temperature_min": "°C", "sunrise": "iso8601", "sunset": "iso8601"}, "daily": {"time": ["2024-09-13"]}, "use": ["2024-09-13T13:22"], "sunset": ["2024-09-14T01:50"]}}
```

**Agent: (thought)** The OpenMeteo tool provided the current weather in Las Vegas.

**Agent:** (final\_answer) The current weather in Las Vegas is 22.3°C with no rain.

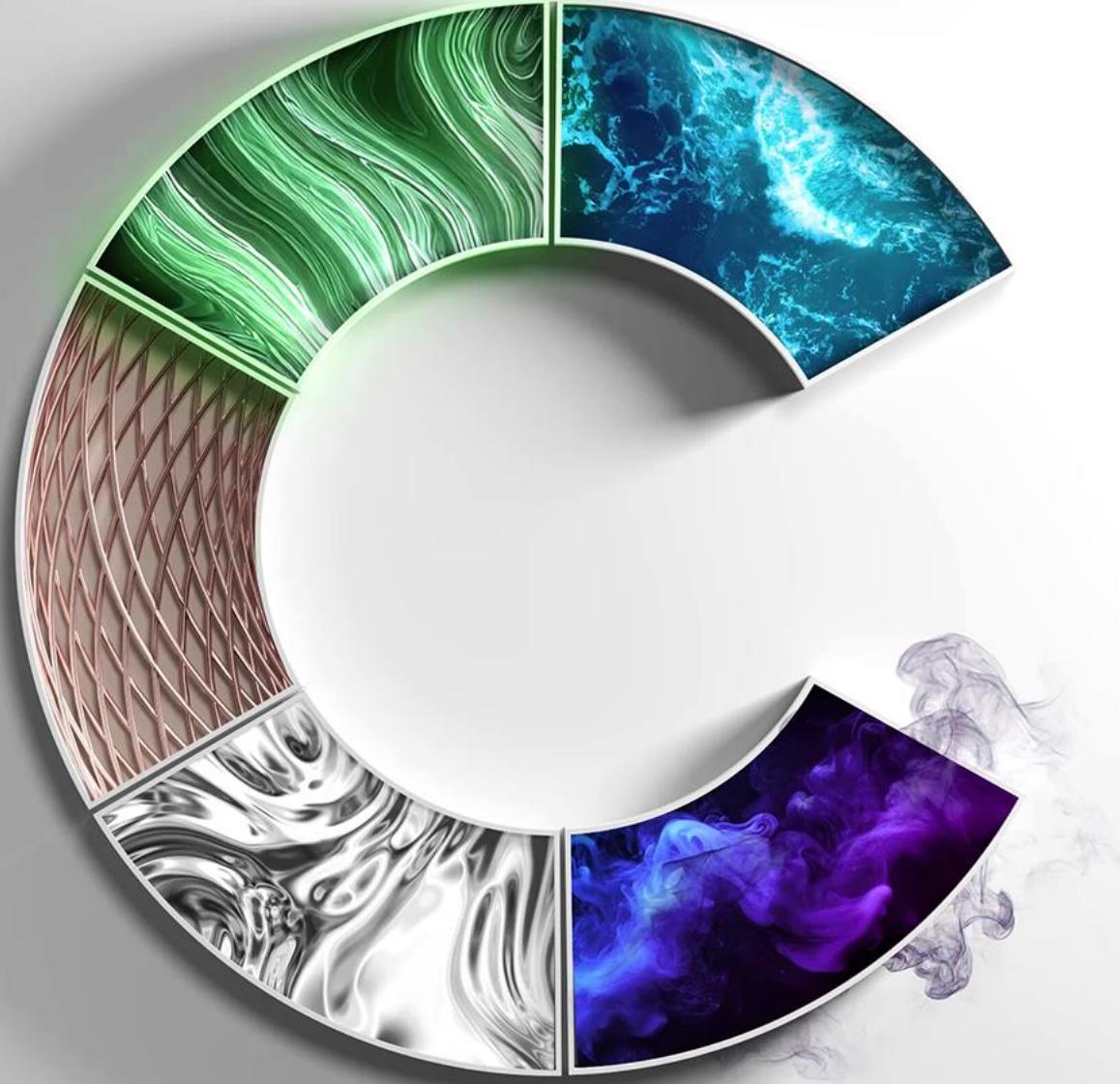
**Agent:** The current weather in Las Vegas is 22.3°C with no rain.



The 2025 CEO Study

# 5 mindshifts to supercharge business growth

Move from productivity to performance with agentic AI



“Pessimists sound smart, but optimists make money. I think there are opportunities when there are market dislocations, and if you feel like you have an information advantage, you should lean into risk.”

[Stephen J. Rich](#)

Chairman, CEO and President  
Mutual of America Financial Group

Change is constant

CEOs are driving their organizations to move faster even as uncertainty presents real challenges

Source: 2025 CEO Study, question E10. To what extent do you agree with the following statements about technology adoption in your organization. 3. "The potential productivity gains from automation are so great that I must accept significant risk to stay competitive." (Strongly) agree. 2025 CEO Study, question E27. Compared to the last two years, how will you drive your organization's pace of transformational change in 2025? Maintain or Accelerate pace. 2025 CEO Study, question E34. To what extent do you agree with the following statements about standards and regulations. 1. "Inconsistent standards and regulations are inhibiting our ability to grow our business." (Strongly) agree.

62%

of CEOs say the potential productivity gains from automation are so great **they must accept significant risk to stay competitive**

and

82%

of CEOs plan to maintain or accelerate their organization's **pace of transformational change** in 2025

but

73%

of CEOs say inconsistent standards and regulations are inhibiting their ability to grow their business

Change is constant

CEOs are prepared to pair human ingenuity with AI agents to supercharge business growth

Source: 2025 CEO Study, question E20. To what extent do you agree with the following statements about agentic AI. 1. "We are actively adopting AI agents and prepared to scale them across our organization." (Strongly) agree. 2025 CEO Study, question E10. To what extent do you agree with the following statements about technology adoption in your organization. 5. "My organization must take advantage of technologies that are changing faster than people can adapt." (Strongly) agree. 2025 CEO Study, question E17. To what extent do you agree with the following statements about generative AI. 3. "I am pushing my organization to adopt generative AI more quickly than some may find comfortable." (Strongly) agree.

but

61%

of CEOs say their organization is *actively adopting AI agents* and is prepared to implement them at scale

65%

of CEOs say their organization must take advantage of technologies that *change faster than people can adapt*

and

53%

of CEOs say they are pushing their organization to *adopt generative AI more quickly* than some people find comfortable

CEOs seek greater clarity and efficiency but identify operational complexity as a critical issue

CEO top priorities  
for the next three years

- #1 Forecast accuracy
- #2 Productivity and profitability
- #3 Product and service innovation

---

CEO greatest challenges  
for the next three years

- #1 Supply chain performance
- #2 Talent recruiting and retention
- #3 Business model innovation

Source: 2025 CEO Study, question E1.1. Which of the following represent your organization's highest priorities over the next 3 years? 2025 CEO Study, question E1.2. Which of the following represent your organization's greatest challenges over the next 3 years?

CEOs must consider five mindshifts to create clarity in crisis

45

Download the full study  
[ibm.co/ceo-2025](https://ibm.co/ceo-2025)



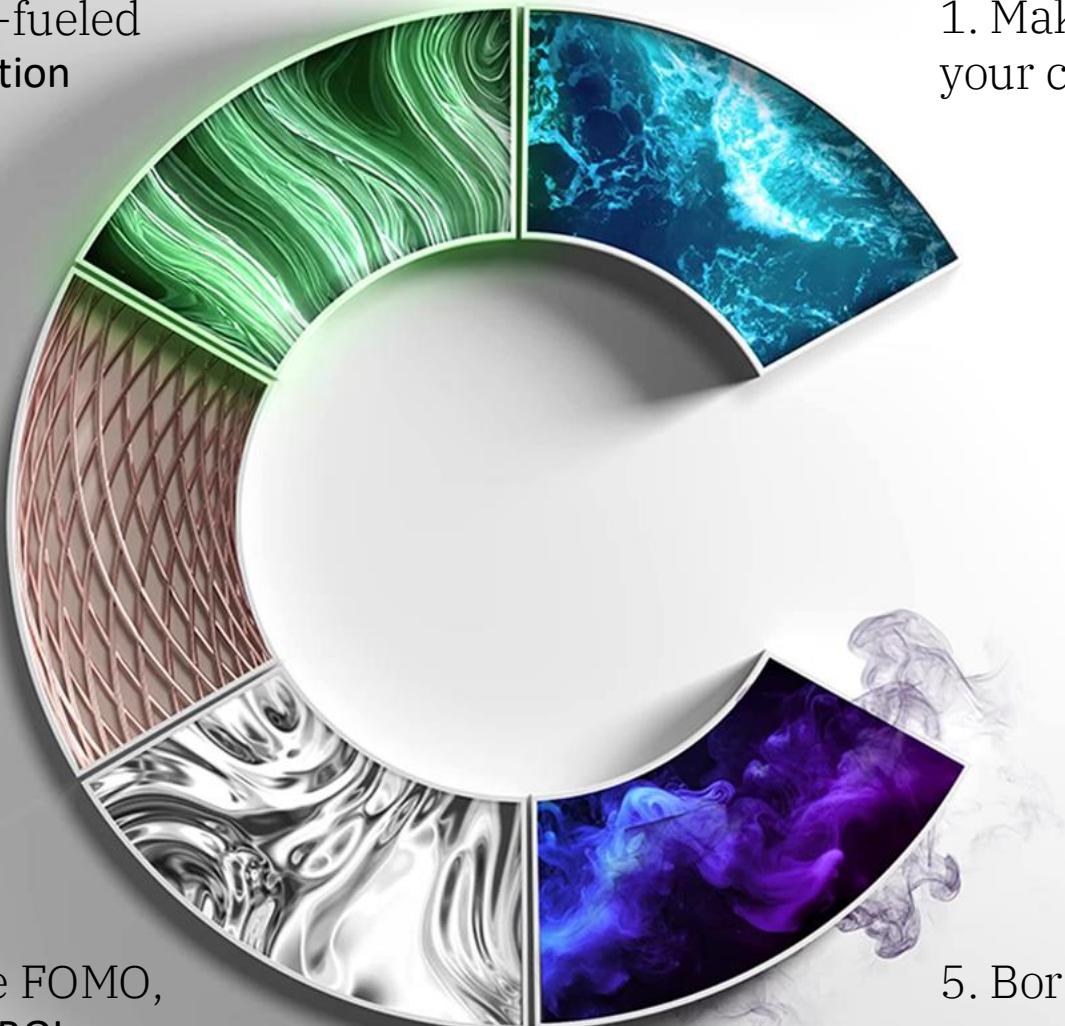
2. Embrace AI-fueled creative destruction

3. Cultivate a vibrant data environment

4. Ignore FOMO, lean into ROI

1. Make courage your core

5. Borrow the talent you can't buy





18TH EDITION

# 2025 TECH TRENDS REPORT

FTSG



Amy Webb Launches 2025 Emerging Tech Trend Report | SXSW LIVE

96K views • Streamed 2 months ago



SXSW

SXSW 2025 Livestream and On Demand Keynotes & Featured Speaker Sessions + VOD Portuguese and Spanish language ...

CC

[https://ftsg.com/wp-content/uploads/2025/03/FTSG\\_2025\\_TR\\_FINAL\\_LINKED.pdf](https://ftsg.com/wp-content/uploads/2025/03/FTSG_2025_TR_FINAL_LINKED.pdf)

[alanbraz.com.br](http://alanbraz.com.br) / [pullrecast.dev](http://pullrecast.dev)

## Future Today Strategy Group's 2025 Tech Trend Report

Our 2025 edition includes 1000 pages, with hundreds of trends published individually in 15 volumes and as one comprehensive report. Download all sections of Future Today Strategy Group's 2025 Tech Trends report at [www.ftsg.com/trends](http://www.ftsg.com/trends).





## 10 Key Takeaways from the FTSG 2025 Tech Trends Report.

**1**  
Living intelligence merges AI, sensors, and biotech into systems that think, adapt, and evolve beyond our grasp.

**2**  
Action models eclipse language models as AI shifts from talking to doing, reshaping automation's frontier.

**3**  
Robots finally break free from factory floors as advanced technology enables real-world adaptability.

**4**  
Agentic AI systems set their own goals and execute complex decisions, augmenting human expertise.

**5**  
Metamaterials rewrite physical limits, as engineered substances transform how we build our world.

**6**  
Tech giants forge unlikely alliances as AI's demands force former rivals to share computing power and data.

**7**  
The climate crisis spurs rapid innovation as extreme weather events accelerate next-gen technology adoption.

**8**  
Nuclear power resurges as AI's energy appetite drives tech giants to invest heavily in small modular reactors.

**9**  
Quantum computing reaches its inflection point as error correction breakthroughs unlock practical use cases.

**10**  
Private enterprise colonizes cislunar space, birthing an economy between Earth and the moon that reshapes commerce.

# AI lacks

**Experience in the  
physical world**

**Human nuance,  
emotional response**

**Physical cause and  
effect**

# FOMO

# FOMA

FTSG



AI: What's bleeding  
edge today might be  
old news... *later today.*

FTSG

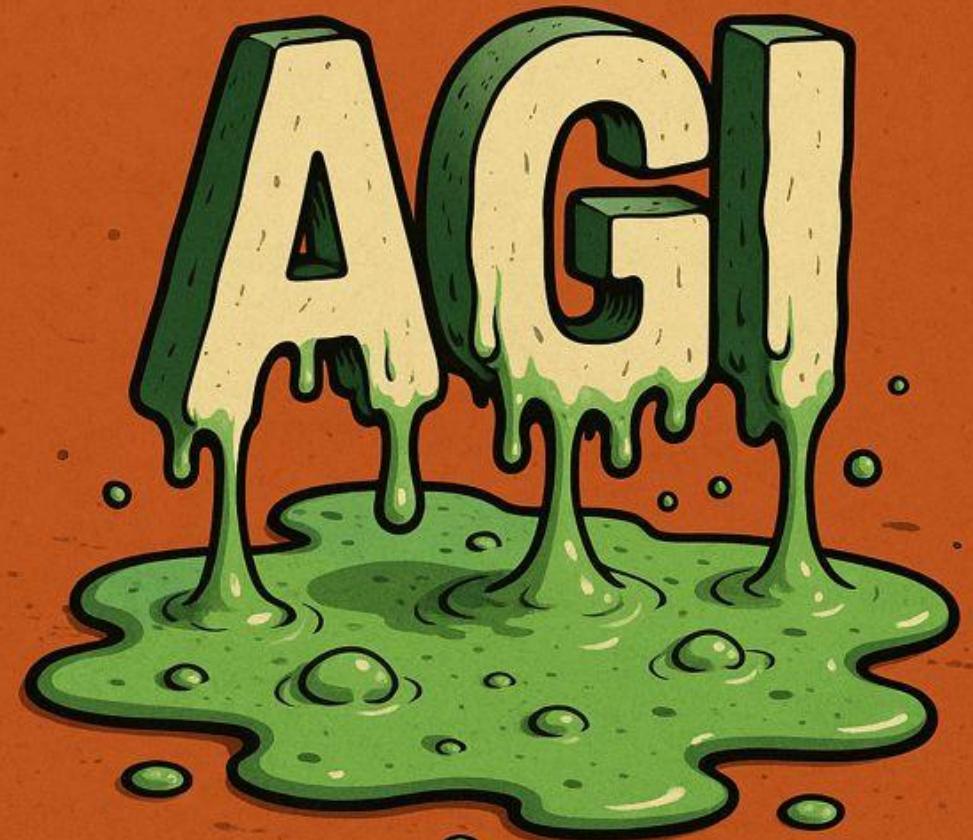
# **“Weeks when decades happen”**

FTSG

[alanbraz.com.br](http://alanbraz.com.br) / [pullrecast.dev](http://pullrecast.dev)

# Mas e o futuro?

[https://www.linkedin.com/posts/ctaurion\\_enquanto-figuras-como-demis-hassab-sam-altman-activity-7330519792300322816-TyLW](https://www.linkedin.com/posts/ctaurion_enquanto-figuras-como-demis-hassab-sam-altman-activity-7330519792300322816-TyLW)





[alanbraz.com.br](http://alanbraz.com.br) / [pullrecast.dev](http://pullrecast.dev)

# Mas e o futuro?



## FUTURE TRENDS



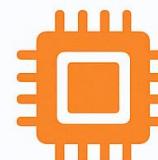
OpenSource  
models



SLMs  
Small Language



Agentic AI



Specific  
hardware

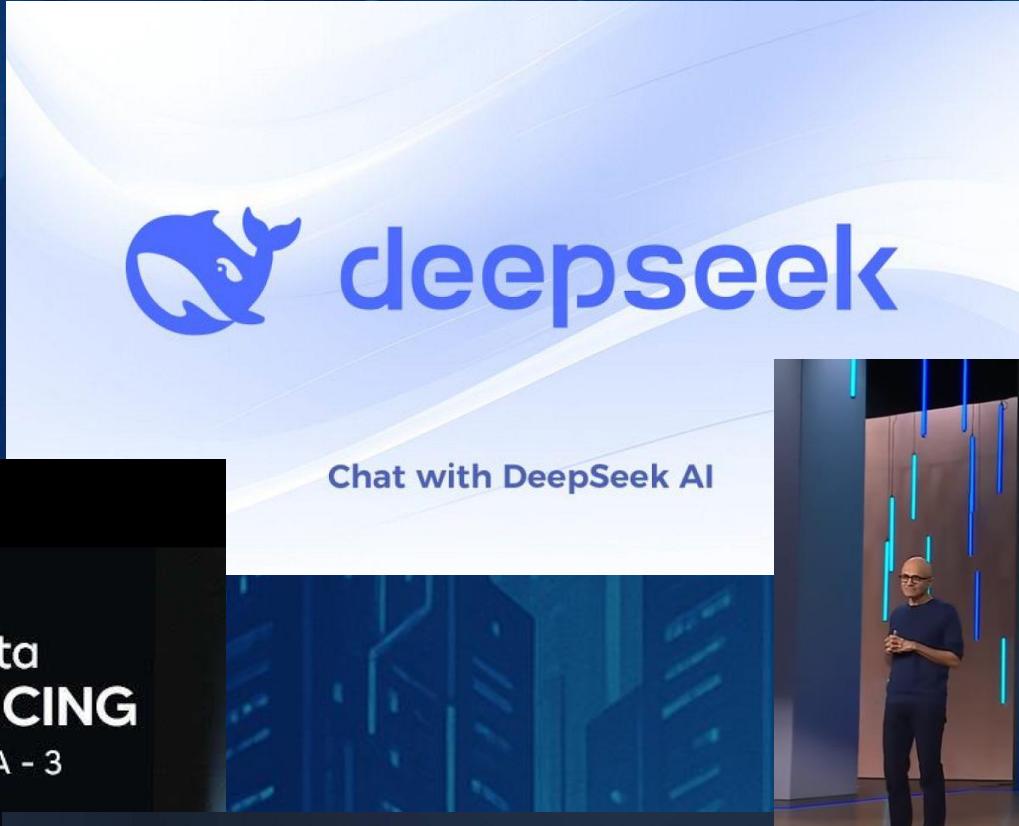


AI Platforms



Quantum  
computing

OpenSource



huggingface.co/deepseek-ai/DeepSeek-R1

Hugging Face Search models, datasets, users...

Models Datasets Spaces Community Docs Pricing

deepseek-ai/DeepSeek-R1 like 12.2k Follow DeepSeek 68.8k

Text Generation Transformers Safetensors deepseek\_v3 conversational custom\_code text-generation-inference fp8 arxiv:2501.12948 License: mit

Model card Files and versions Community 216 Edit model card

Downloads last month 735,368

Safetensors Model size 685B params Tensor type BF16 · F8\_E4M3 · F32 Chat template Files info

Inference Providers NEW

Hyperbolic Examples

Text Generation

Input a message to start chatting with deepseek-ai/DeepSeek-R1.

Model size 685B params

DeepSeek-R1

DeepSeek Homepage Chat DeepSeek R1 Hugging Face DeepSeek AI Discord DeepSeek AI WeChat DeepSeek AI Twitter deepseek ai License MIT

Paper Link

# SLM

Quantization  
Distillation  
<20b models

LARGE LANGUAGE MODEL	Parameters
Phi-1.5	1.3B
Phi-2	2.7B
Llama2	7B, 13B, or 70B
BloombergGPT	50B
Claude2	130B
GPT-3	175B
GPT-4 "32k"	1.76T

<https://kelvin.legal/understanding-large-language-models-what-are-paramters/>

[Discord](#) [GitHub](#) [Models](#) Search models[Sign in](#)[Download](#)

## deepseek-r1

`ollama run deepseek-r1`

⬇️ 45.5M Downloads ⏲ Updated 3 months ago

DeepSeek's first-generation of reasoning models with comparable performance to OpenAI-o1, including six dense models distilled from DeepSeek-R1 based on Llama and Qwen.

[1.5b](#) [7b](#) [8b](#) [14b](#) [32b](#) [70b](#) [671b](#)

### Models

[View all →](#)

Name	Size	Context	Input
deepseek-r1:latest	4.7GB	128K	Text
deepseek-r1:1.5b	1.1GB	128K	Text
deepseek-r1:7b <span>latest</span>	4.7GB	128K	Text
deepseek-r1:8b	4.9GB	128K	Text
deepseek-r1:14b	9.0GB	128K	Text
deepseek-r1:32b	20GB	128K	Text
deepseek-r1:70b	43GB	128K	Text
deepseek-r1:671b	404GB	160K	Text



New chat

Today

H100 GPU Requirements for ...

7 Days

Explicação simples de inteligênci

2025-04

Como remover linhas duplicadas

Alan Braz's Professional Profile St

Currículo de Pós-Graduação em C

2025-01

Meaning and Explanation of GSH

NBA Games in NY or Brooklyn Fe

Cardápio semanal saudável para f



## How Many H100 GPUs Are Needed for a 671B Model?

### 1. Memory Requirements (FP16/BF16)

- **Parameters:**  $671 \text{ billion} \times 2 \text{ bytes} = 1,342 \text{ GB (1.34 TB)}$  for weights alone.
- **Additional memory for:**
  - **Activations** (depends on sequence length & batch size)
  - **Optimizer states** (if training, Adam adds ~2x model size)
  - **KV Cache** (for inference)

### 2. GPU Memory per H100

- **H100 80GB PCIe/SXM:** 80GB per GPU.
- **H100 with NVLink + FP8** could help but assume **FP16** for safety.

### 3. Inference vs. Training Breakdown

#### Inference (Lower Memory Needs)

- **Model weights alone:** 1,342 GB →  $\lceil 1,342 / 80 \rceil = 17 \text{ GPUs (bare minimum)}$
- **With activations & KV cache,** likely 20–24 H100s for reasonable batch sizes.

#### Training (Much Higher Memory Needs)

- **Full training memory** (weights + gradients + optimizer states):
  - **Weights:** 1,342 GB
  - **Gradients:** 1,342 GB

Message DeepSeek

DeepThink (R1)

Search

how many H100 GPUs I need to run deepseek r1 671b?

## 4. Estimated GPU Count for 671B Model

Scenario	Minimum H100 GPUs	Recommended GPUs
<b>Inference</b>	17	20–24+
<b>Training</b>	64	128+ (w/ ZeRO-3)

## Key Considerations

### 1. Memory Optimization:

- **Use FP8 quantization** (cuts memory in half vs. FP16).
- **FlashAttention-3** for efficient KV cache.
- **Model sharding** (Tensor/Pipeline Parallelism).

### 2. Training Practicality:

- Training a **671B model** is **extremely expensive** (~millions in compute).
- **Mixture of Experts (MoE)** could reduce active parameters.

### 3. Inference Bottlenecks:

- Even with 24 GPUs, latency may be high without optimizations like **continuous batching**.

## Final Answer

- **For inference:** **20–24 H100 (80GB) GPUs** (with optimizations).
- **For training:** **128+ H100s** (likely requiring a multi-node cluster).



O que você procura?



Faça seu login



TODOS OS DEPARTAMENTOS



COMPUTADORES



ARMAZENAMENTO



NAS



NOTEBOOKS



HARDWARE



PLACAS DE VÍDEO



PROCESSADORES



LANÇAMENTOS



OFERTA DO DIA

Waz / Hardware / Placa De Vídeo (VGA)

## Placa de vídeo NVIDIA H100 (80GB, PCIe 5.0) - PNY NVH100TCGPU-KIT

Placa de vídeo com GPU NVIDIA H100 (1,095GHz), 80GB de memória HBM2e (5.120bits), interface PCI Express 16x v5.0.



Cod.129603

PNY



Vendido e entregue por

POR: R\$ 249.999,99

no pix já com 15% de desconto

OU R\$ 294.117,64

COMPRAR

Pré venda

Opções de parcelamento

1x de R\$ 276.470,58 com 6% de desconto

2x de R\$ 139.705,88 com 5% de desconto

[Ver mais](#)

3x de R\$ 93.627,45 com 4,5% de desconto

4x de R\$ 70.588,23 com 4% de desconto

Calcular o Frete

Digite seu CEP

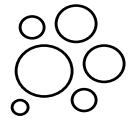
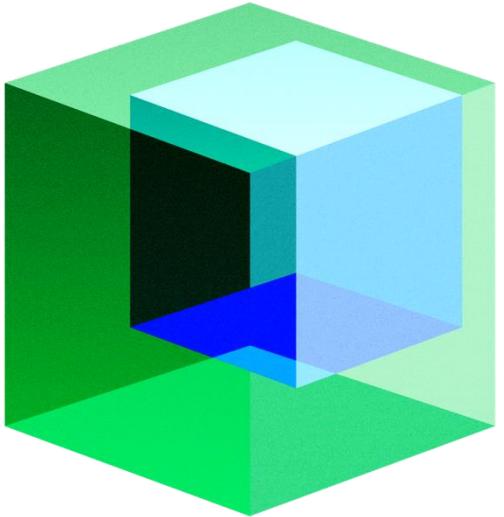
**CALCULAR**

[Não sei meu CEP](#)

Preços e formas de pagamento válidos apenas para o website. Consulte condições nas lojas.

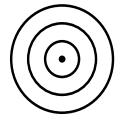
# IBM Granite

Uma família de modelos de IA abertos, de alto desempenho e confiáveis para acelerar a adoção da IA corporativa



## Aberto

- Código aberto no **Apache 2.0**
- **Transparência** dos dados e métodos de treinamento
- **Personalize** com seus dados



## Eficiente

- Gama diversificada de modelos **adequados à finalidade**
- Projetado para **escalabilidade**
- Recursos de **raciocínio** que mantêm o **desempenho** geral de forma exclusiva
- Recursos de **visão** otimizados para compreensão de documentos corporativos



## Confiável

- Identificação de IP
- IA **responsável** e segura
- Guardrails para **proteger dados e mitigar riscos**
- Abordagem de raciocínio que preserva exclusivamente a **segurança** do modelo

# Getting started with Granite models

[ibm.com/watsonx/developer](https://ibm.com/watsonx/developer)

[huggingface.com/ibm-granite](https://huggingface.com/ibm-granite)



## Vision model

- Granite-Vision-3.2-2B



## Language models

- Granite-7B-Base
- Granite-3.1-1B-A400M
- Granite-3.1-3B-A800M
- Granite-3.1-2B
- Granite-3.1-8B
- Granite-3.2-2B
- Granite-3.2-8B



## Guardian models

- Granite-Guardian-HAP-125M
- Granite-Guardian-HAP-38M
- Granite-Guardian-3.1-2B
- Granite-Guardian-3.1-8B
- Granite-Guardian-3.2-8B
- Granite-Guardian-3.2-5B
- Granite-Guardian-3.2-3B-A800M



## Accelerator models

- Granite-3B-Code-Instruct-Accelerator
- Granite-8B-Code-Instruct-Accelerator
- Granite-20B-Code-Instruct-Accelerator
- Granite-34B-Code-Instruct-Accelerator
- Granite-7B-Instruct-Accelerator
- Granite-3.0-8B-Instruct-Accelerator



## Code models

- Granite-34B-Code
- Granite-20B-Code
- Granite-8B-Code
- Granite-3B-Code



## Time Series models

- Granite-TimeSeries-TTM-r1
- Granite-TimeSeries-PatchTSMixer
- Granite-TimeSeries-PatchTST
- Granite-TimeSeries-TTM-r2
- Granite-TimeSeries-TTM-r2.1



## Geospatial models

- Earth
- Granite-EarthObservation-HLS-Biomass
- Granite-EarthObservation-HLS-CanopyHeight
- Granite-EarthObservation-HLS-Landslide
- Weather and climate
- Granite-WeatherClimate-Precip-Downscaling
- Granite-WeatherClimate-WindForecasting

- Hostinger  
2 CPU  
R\$40/mês
- OpenWebUI
- Ollama

The screenshot shows a web-based AI chat interface. The URL in the address bar is [chat.alanbraz.com.br/c/d267d8d8-2979-4cf9-af99-e14a2ea14e9b](https://chat.alanbraz.com.br/c/d267d8d8-2979-4cf9-af99-e14a2ea14e9b). The interface has a dark theme with a sidebar on the left containing various AI-generated responses and a main chat area on the right.

**Left Sidebar (Chats):**

- New Chat
- Notes
- Workspace
- Search
- Chats
  - Previous 7 days
  - AI Explained for an 89-Year-... (highlighted)
  - Assistência e Comunicação
  - Joke Tell
  - Cat chased fox
  - Joke Response Shared
  - Joke Response
  - New Chat
  - Previous 30 days
  - Descrição de Página
  - Greeting, Olá!
  - Greeting, Hello
  - New Chat
  - New Chat
  - Vocabulary Practice
  - New Chat
  - New Chat

**Right Main Area:**

A message from the AI model "granite3-moe:1b" is displayed:

Como você explicaria inteligência artificial brevemente para minha avó de 89 anos?

**AI Response (granite3-moe:1b):**

A inteligência artificial é uma habilidade de computadores que permite que eles possam entender, processar e gerenciar informações de maneira mais eficiente. Eles podem aprender a partir de dados e usar essas informações para realizar tarefas, como enviar mensagens, jogar jogos ou resolver problemas complexos. Essas ferramentas são capazes de ser usadas em uma variedade de aplicativos, incluindo o transporte automático, a assistência aos seniores e a gestão de processos.

Em relação à sua idade, o melhor aconselhamento é fornecer informações detalhadas e cuidado, para ajudar a entender bem como funciona a inteligência artificial. É importante lembrar que essas ferramentas não possuem sentimentos nem consciência, portanto a sua confiança deve ser baseada na verificação completa dos recursos disponíveis e na sua experiência com o sistema atual.

Sua idade também pode influenciar a manutenção necessária para garantir que as ferramentas sejam delicadas e eficazes, pois elas podem requerer atualizações muitas vezes dependendo da quantidade de dados processados. Portanto, é importante ter em conta uma política de manutenção constante para prolongar a vida útil das ferramentas.

**Bottom Input Area:**

Send a Message

+ Web Search Code Interpreter

AB Alan Braz

# Agentic AI

[https://www.youtube.com/watch?  
v=EDb37y\\_MhRw](https://www.youtube.com/watch?v=EDb37y_MhRw)





## AI Agents

Streamline repetitive tasks



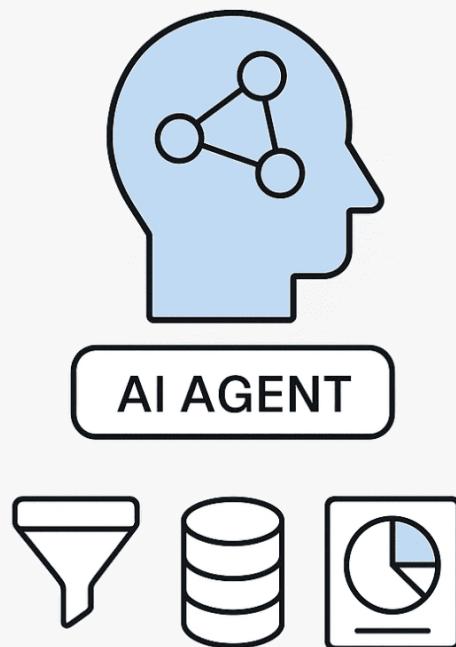
## Agentic AI

Manage complex projects

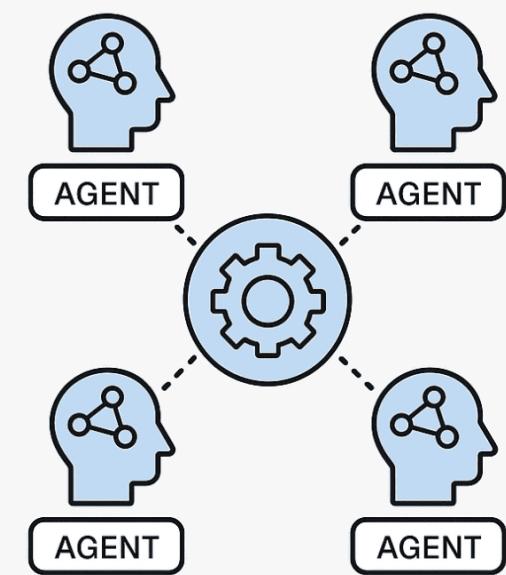
<https://www.mygreatlearning.com/blog/ai-agents-and-agentic-ai/>

<https://theaiinsider.tech/2025/05/19/whats-the-difference-between-ai-agents-and-agentic-ai-new-study-separates-signal-from-noise-in-the-ai-agent-boom/>

## AI Agents



## Agentic AI



# A estrada até aqui...



## Chatbots

- Baseado em regras
- Determinístico
- Pré-programado



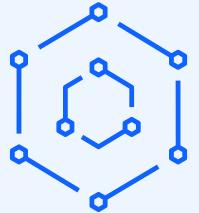
## Assistentes virtuais

- NLU
- Reconhecimento de intenções
- Técnicas de ML e DL



## Assistentes de IA

- Generative AI e LLMs
- Tarefas prescritivas
- Processos únicos

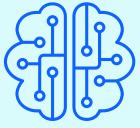


## Agentes de IA

- Multi processos
- Tomando ações de forma autônoma
- Auto-correção

# Capacidades da IA estão crescendo rapidamente

IA que cria



*Generative AI*

IA que prediz



*Machine learning*

IA que conversa



*AI assistants*

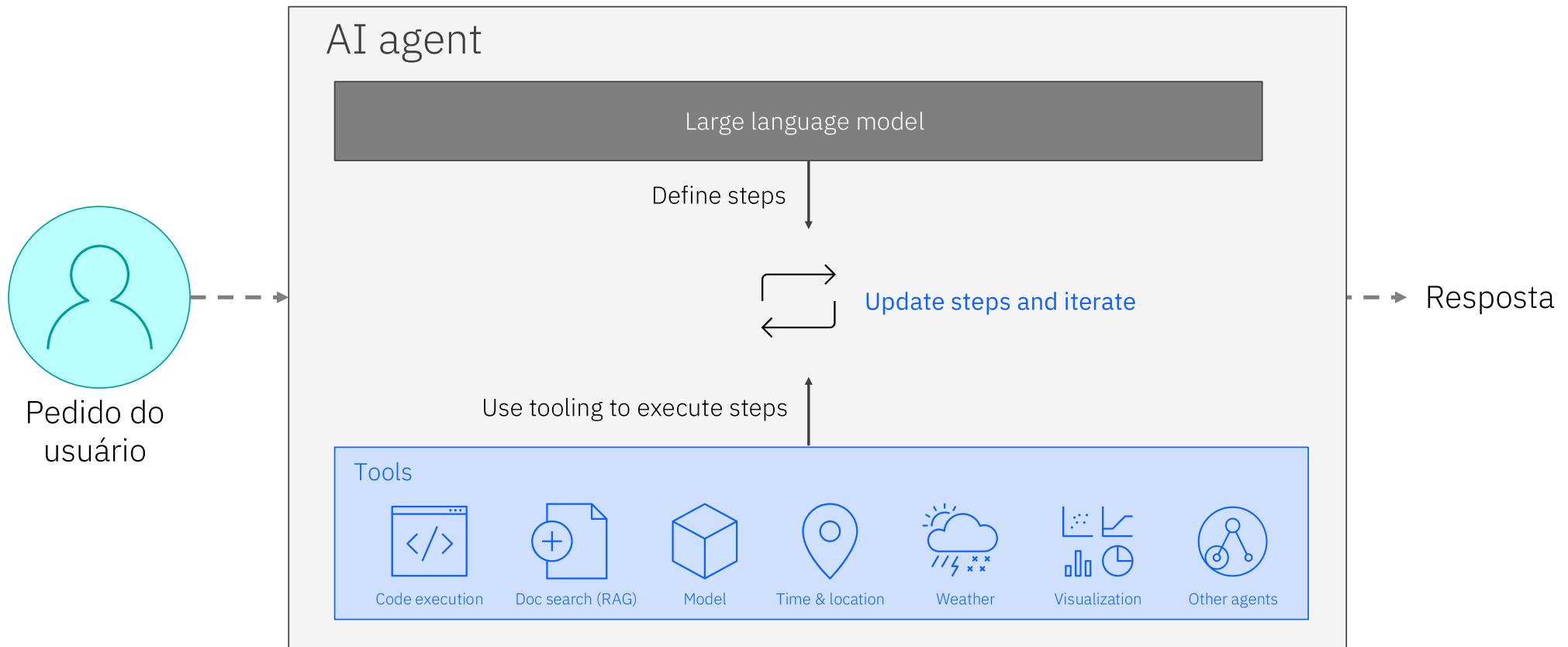
IA que trabalha



*AI agents*

# O que são AI Agents e como eles trabalham?

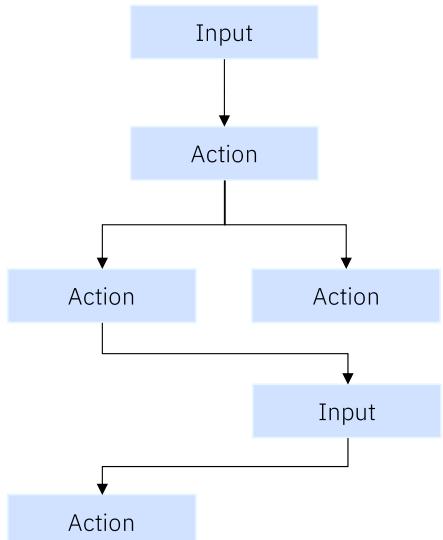
Um AI agent é um programa cuja lógica de execução é controlada por LLM



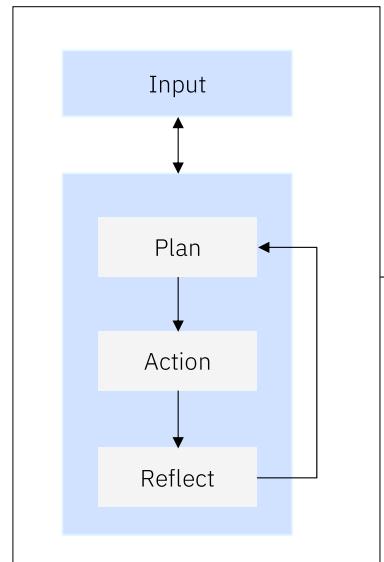
# Evolução

Traditional assistants → Single-agent assistants → Multi-agent assistants

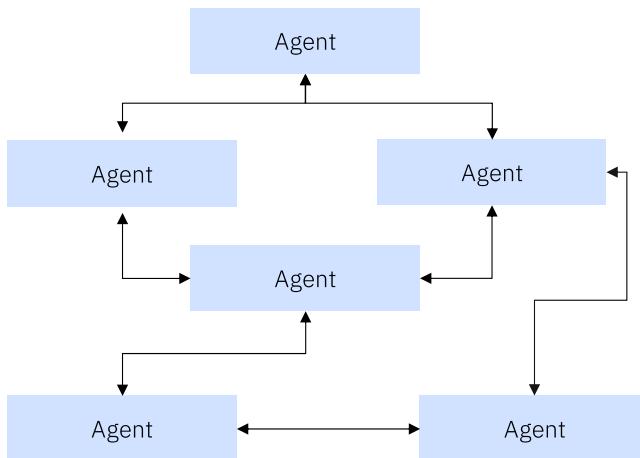
- Rule based (if x, do y)
- Predefined action paths



- Task based (e.g., flight booker)
- Performance constraints as scope of task increases
- Limited control



- Domain based (e.g., travel agent)
- Specialized agents (planner, flight booker, hotel booker, etc.) work together, improving system performance
- Control over how agents communicate
- Multiple architecture options (network, hierarchical, supervisor, custom)



# Estratégia IBM: AI Agents para negócios



AI Agents  
para negócios

CONVERSACIONAL | ORQUESTRA HABILIDADES | CONTEXTUALIZADO  
OMNI-CHANNEL | MULTI-CLOUD

IA GENERATIVA

watsonx™

CLASSIFICAR | GERAR | SUMARIZAR | EXTRAIR

CONHECIMENTO

BUSCA SEMANTICA | VECTOR DB

AUTOMAÇÃO

RPA | WORKFLOW | DECISION | SISTEMAS EXISTENTES

INTEGRAÇÕES

genesys

NICE

CXone

Marketo™  
An Adobe Company

twilio

Microsoft  
Salessoft

coupa  
servicenow

box  
elastic

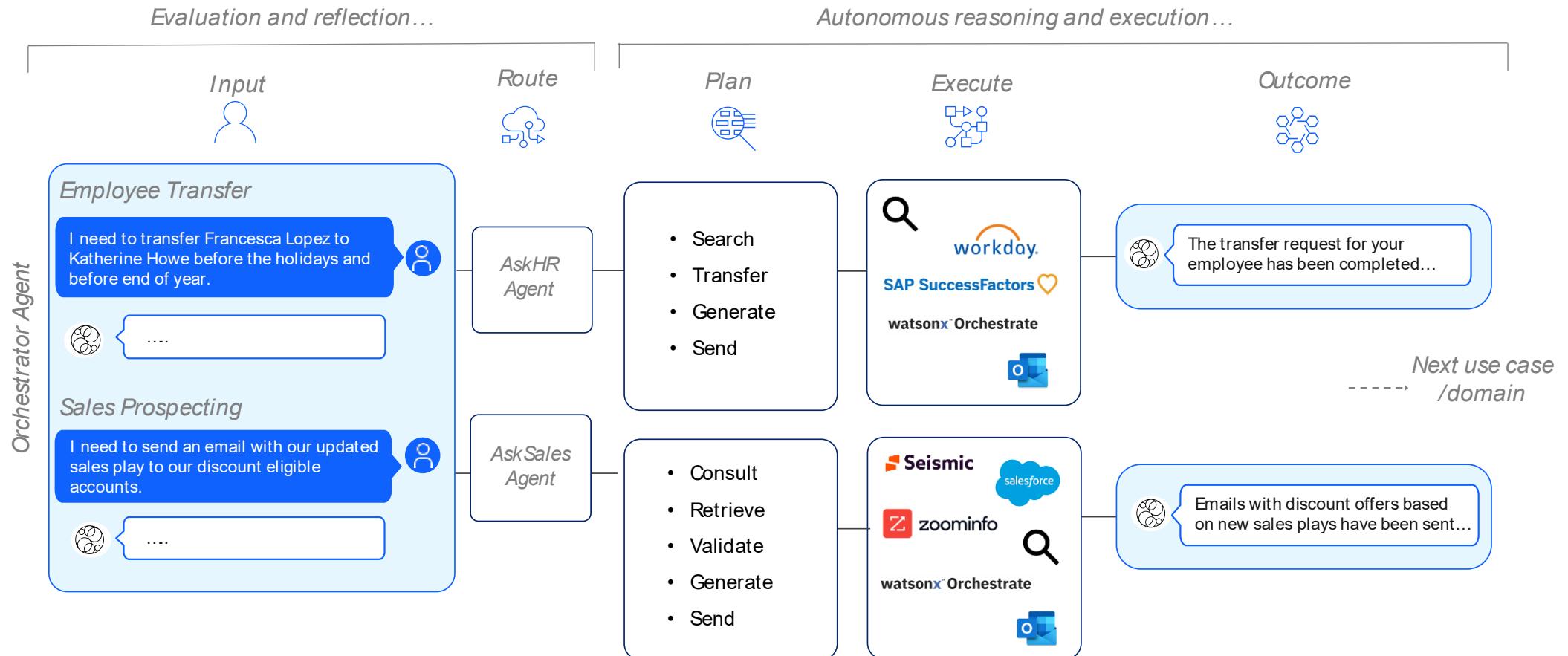
dun&bradstreet  
SAP Ariba

IBM  
Planning Analytics

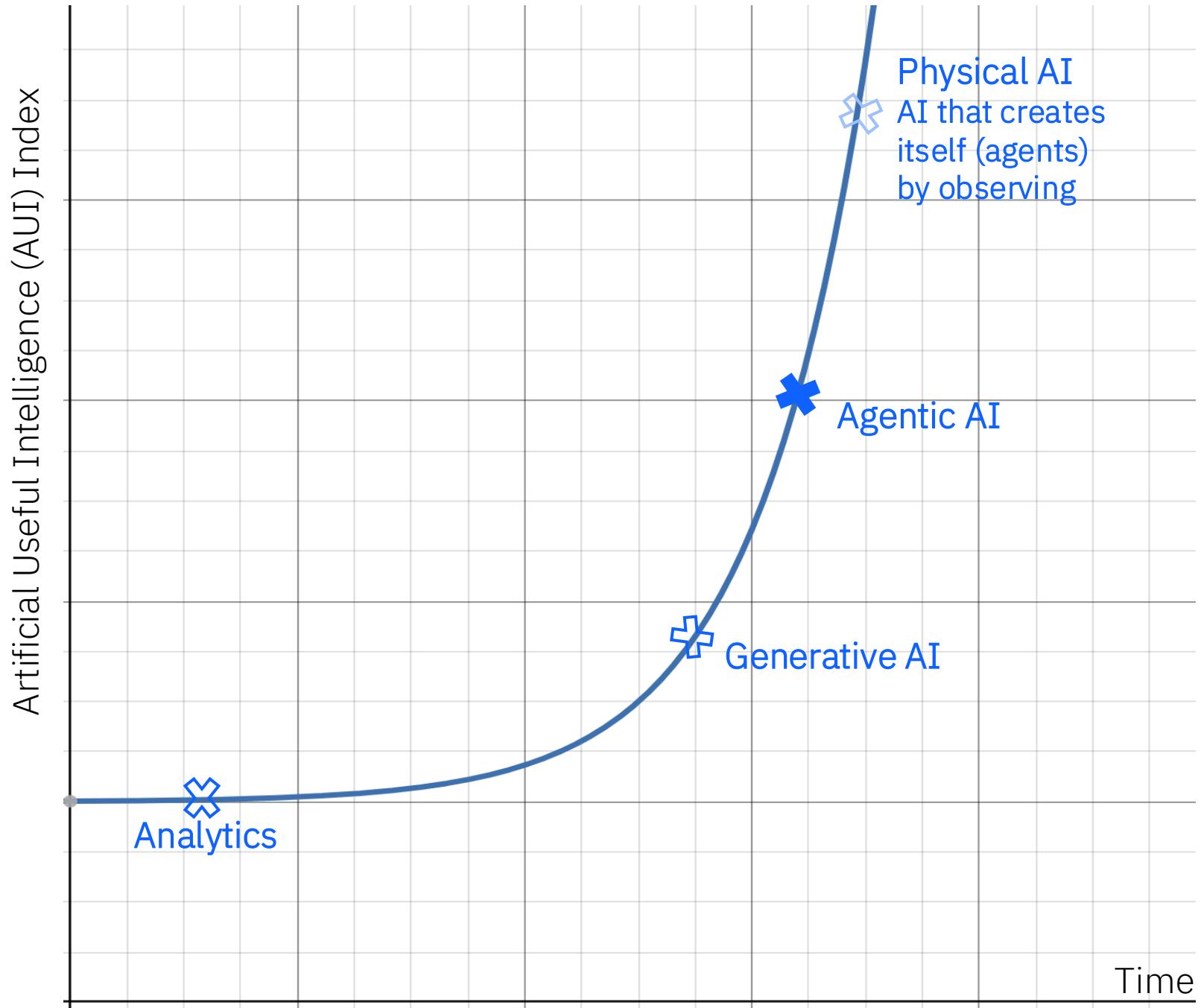
ORACLE  
thisway

SAP SuccessFactors  
SAP

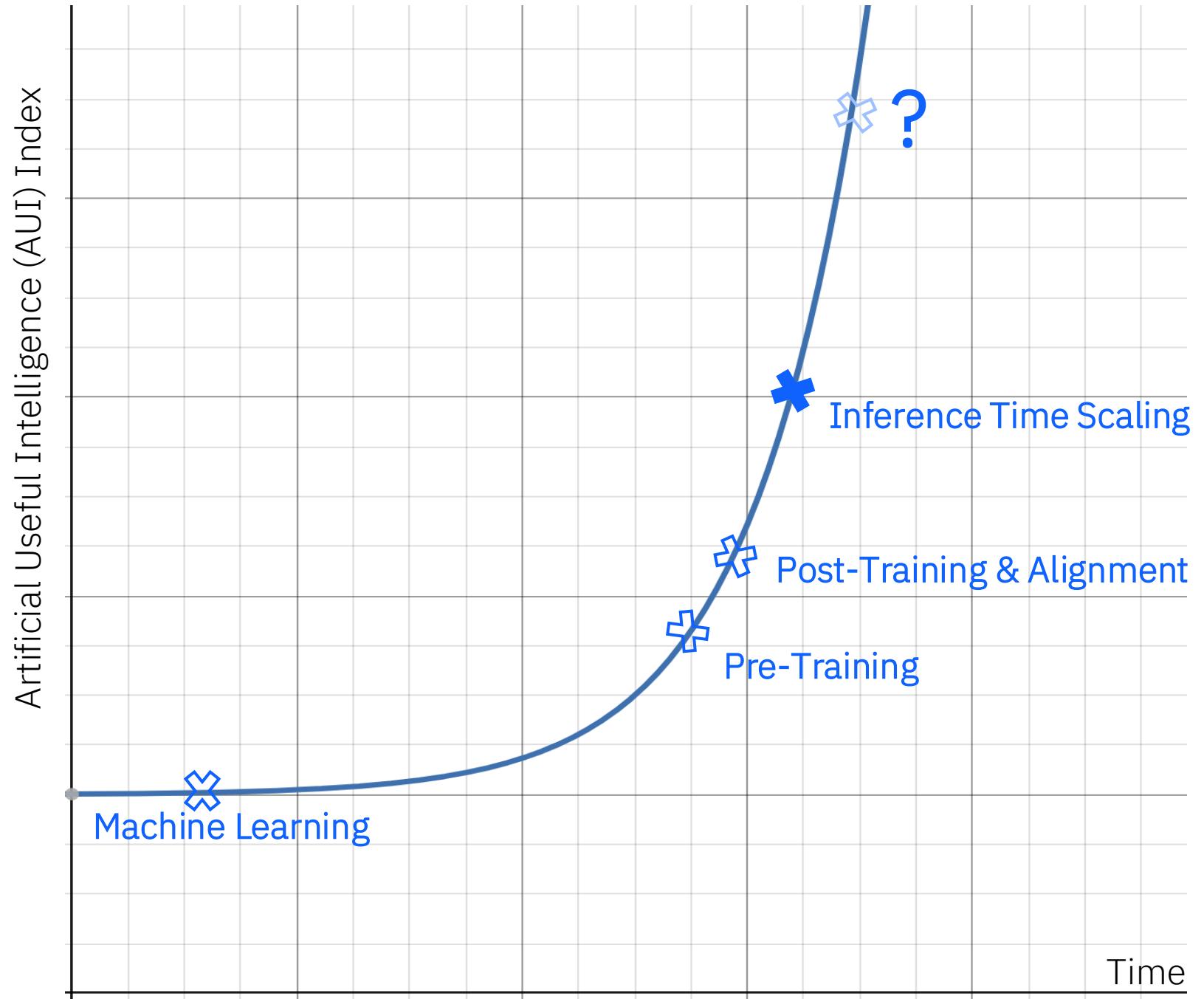
# Orchestrator Agent



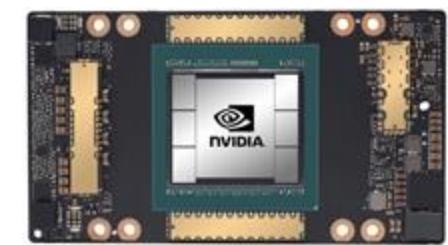
# Artificial Intelligence Futures



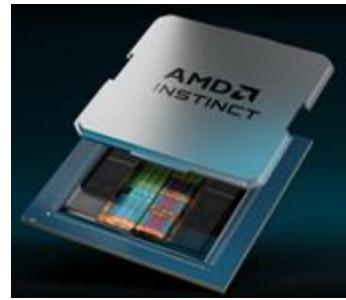
# AI Technology Futures



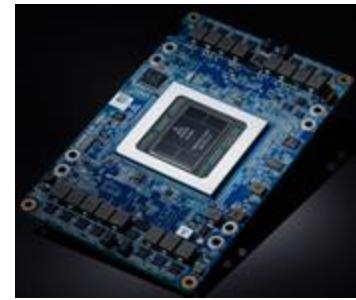
# Hardware Support



NVIDIA GPU



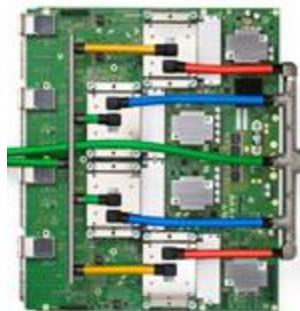
AMD GPU



AWS Neuron



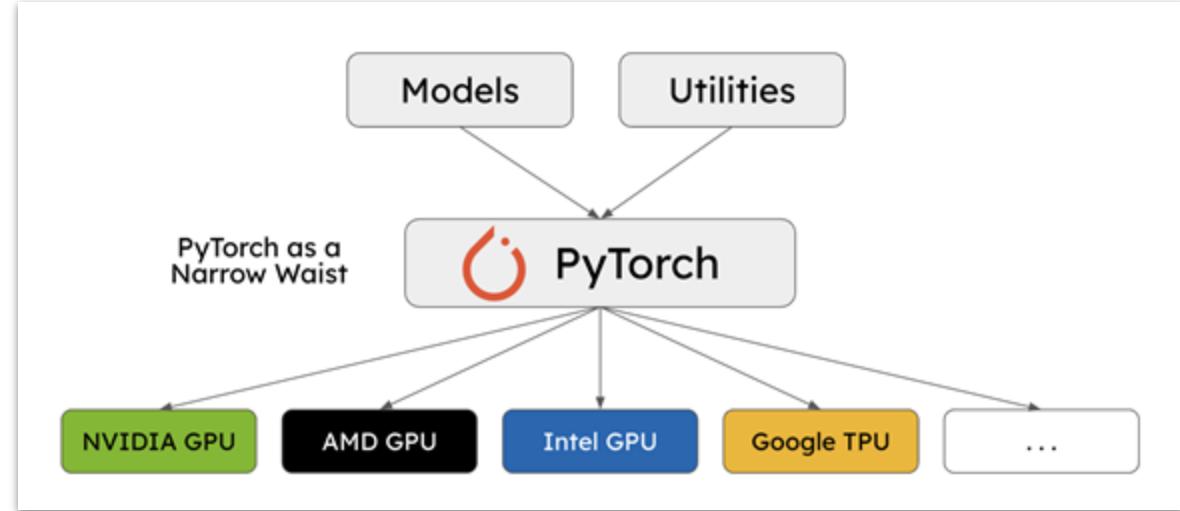
Intel CPU/GPU



Google TPU



Intel Gaudi



IBM Spyre

# Plataformas



Microsoft Azure



Google AI

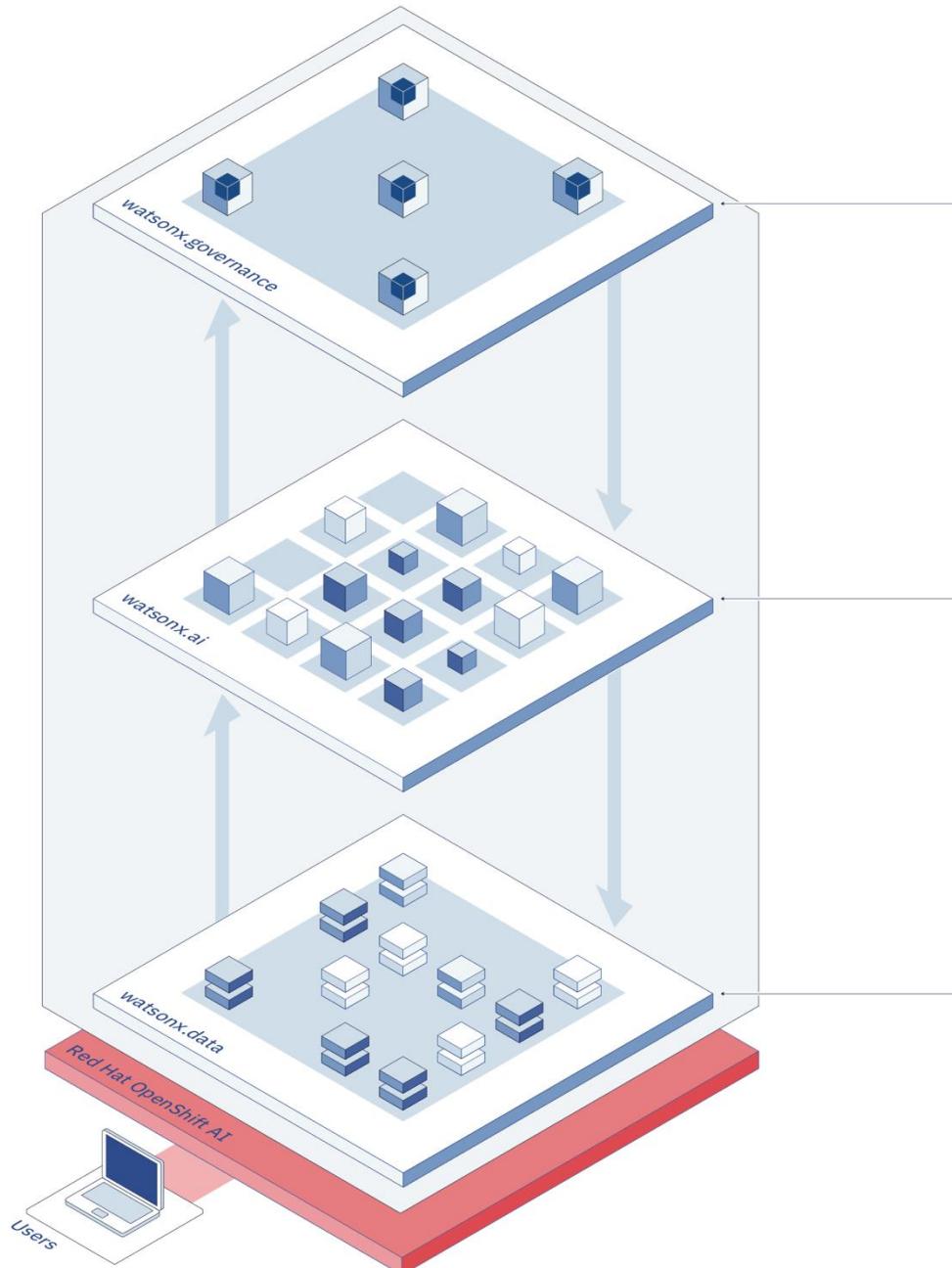


OpenAI



NVIDIA

# watsonx, the platform for AI value creators



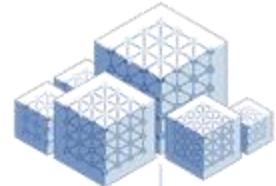
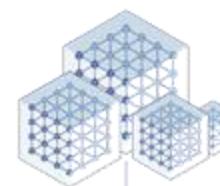
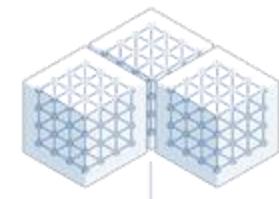
## watsonx.governance

End-to-end toolkit encompassing both data and AI governance to enable responsible, transparent, and explainable AI workflows.



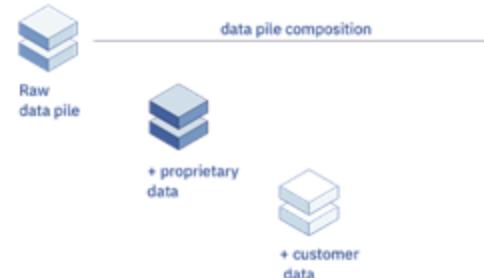
## watsonx.ai

A next generation enterprise studio for AI builders to train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables you to build AI applications in a fraction of the time with a fraction of the data.



## watsonx.data

Fit-for-purpose data store optimized for governed data and AI workloads, supported by querying, governance, and open data formats to access and share data.





## Trechos



Computador quântico: como funciona a tecnologia que promete revolucionar a...



Fantástico apresenta clipe especial para celebrar os 60 anos da TV Globo



2025: confira a nova abertura do Fantástico



As origens e a formação religiosa do muito portenho Jorge Bergoglio

Fantástico

### Computador quântico: como funciona a tecnologia que promete revolucionar a vida na Terra

14 min

CC

Empresas e países investem pesado para dominar ciência que pode ajudar a curar doenças, combater as mudanças climáticas e expandir o conhecimento sobre o universo.

<https://globoplay.globo.com/v/13550697/>

alanbraz.com.br / pullrecast.dev



# Ansiedade?!?!

Não se trata de prever o futuro exato, mas de planejar de forma inteligente e evitar erros dispendiosos. Lidere a transformação com coragem, impulsionado por dados, capacitando pessoas e parceiros, e sempre com foco em ética e governança.