

# Agentes de IA e o Futuro da tecnologia

Alan Braz

Lead Research Developer - IBM

[alan@pullrecast.dev](mailto:alan@pullrecast.dev)



# Alan Braz

Mestre em Ciência da Computação pela Unicamp, com ênfase em Engenharia de Software Ágil.

Líder de Desenvolvimento na IBM Research (Pesquisa) atuando na transformação de projetos de pesquisa para produtos em nuvem híbrida, liderando equipes e automatizando o processo de desenvolvimento, visando ofertas open-source ou novos produtos como o IBM watsonx.ai.

Criador do podcast PullreCast falando sobre o impacto da Inteligência Artificial nas carreiras de tecnologia e nos negócios!

Conselheiro de IA do Centro de Referência em Inovação da Fundação Dom Cabral.

Mentor, palestrante e professor de IA na Alura, CNEX e FDC.



# IA: a oportunidade de uma década

16T

A IA é a **maior oportunidade econômica** da nossa era, estimando-se que contribua com 16 trilhões de dólares para o PIB global até 2030

45%

de ganhos de produtividade

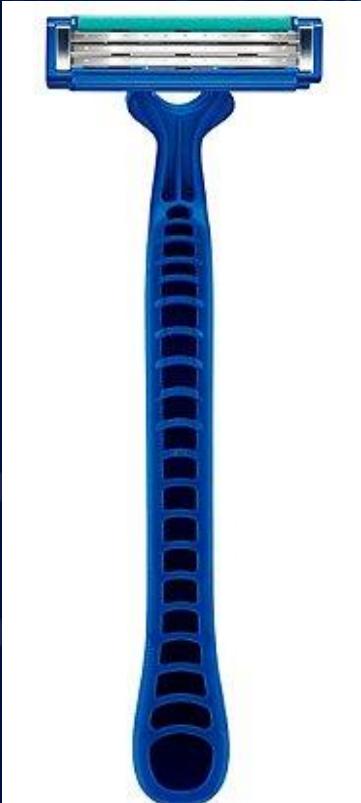
80%

das empresas terão incorporado **Gen AI** em seus processos de negócios em 2026





# Sinônimos



# ChatGPT



## Examples

"Explain quantum computing in simple terms" →

"Get any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



## Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



## Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and current affairs

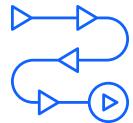
# O melhor da IA

Quando usar

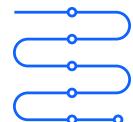
## Recursos de IA **Tradicional**



Preditivo/Prescritivo



IA conversacional dirigida



IA de Visão Dirigida

Quando usar

## Recursos de IA **Generativa**



Sumarização



Pesquisa conversacional



Criação de conteúdo

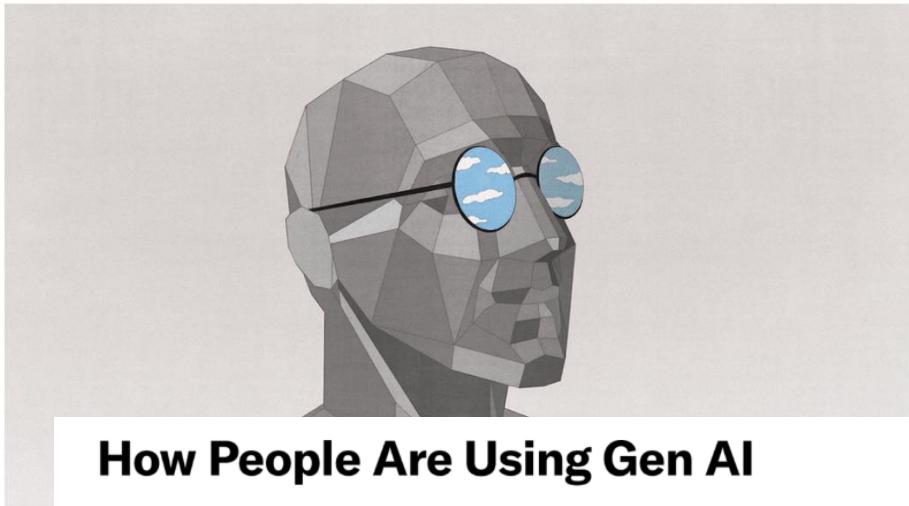


Criação de código

# How People Are Really Using Gen AI in 2025

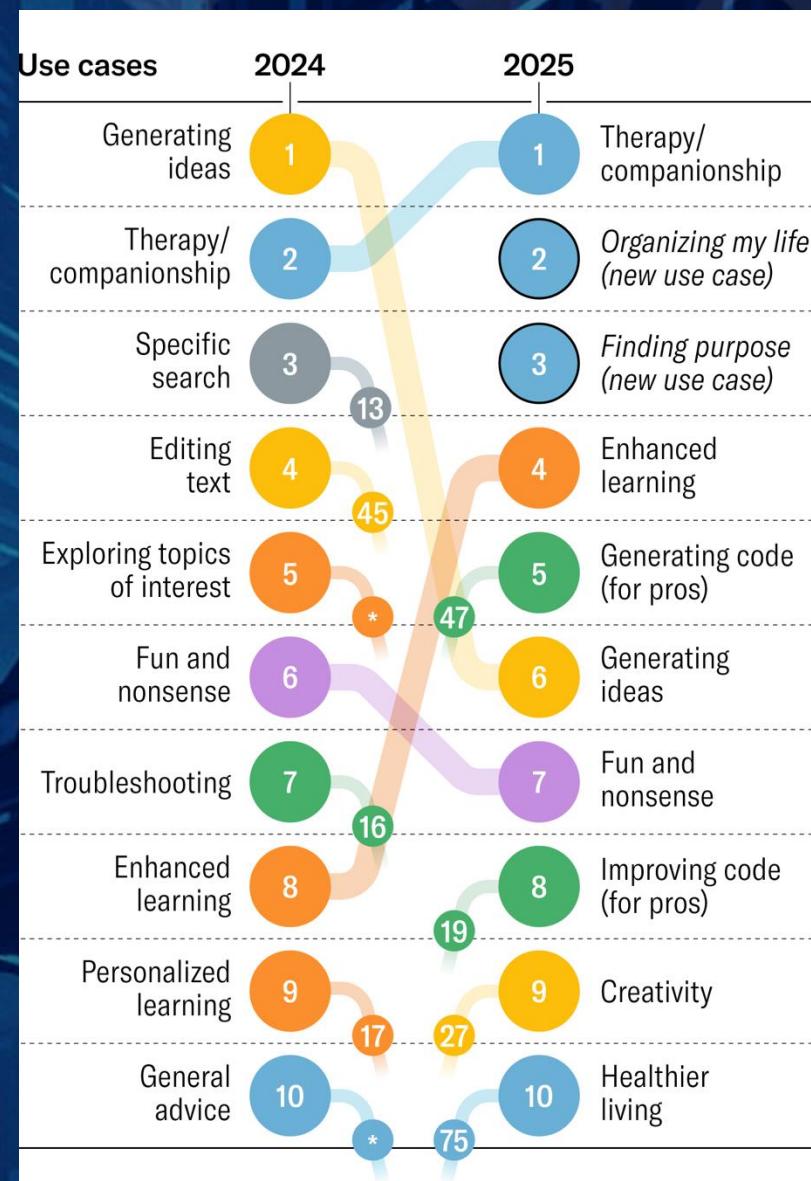
by Marc Zao-Sanders

April 9, 2025



## How People Are Using Gen AI

Have people found ways for generative AI to help lighten their workloads, increase their productivity, or think through problems in new ways? To understand how individuals are using the technology, researchers mined web forums like Quora and Reddit, filtering through tens of thousands of posts to identify 100 different use cases, which they then organized into six themes.



# Terapia por chatbot e amigos virtuais: 10 principais usos da IA em 2025

Principais buscas em IA sofreram uma mudança de aplicações técnicas para aplicações emocionais

Mariana Valbão, da CNN

30/05/25 às 19:15 | Atualizado 30/05/25 às 19:33

- 
- 1.** Terapia e Companheirismo
  - 2.** Organizando minha vida
  - 3.** Encontrando um propósito
  - 4.** Aprendizagem aprimorada
  - 5.** Gerando código
  - 6.** Gerando ideias
  - 7.** Diversão e bobagens
  - 8.** Aprimorando o código
  - 9.** Criatividade
  - 10.** Vida mais saudável

<https://www.cnnbrasil.com.br/tecnologia/terapia-por-chatbot-e-amigos-virtuais-10-principais-usos-da-ia-em-2025/>

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

# GENERATIVE AI USAGE BY GENERATION

Top GenAI Uses	Behavior/Approach
Boomers  Productivity, health, chatbots	Cautious adopters, utility-focused
Gen X  Work tasks, parenting, finances	Practical, work-life balance
Millennials  Content creation, career growth side hustles	Creative, self-empowered
Gen Z  Social media, AI friends, study tools	Expressive, experimental, mobile-first
Gen Alpha  Education, AI games voice interaction	Native to AI, highly visual & interactive

TECH

## 69% dos membros da Geração Z pedem coisas ao ChatGPT educadamente com uma intenção oculta

Um em cada 10 jovens da Gen Z gostaria que seu chefe fosse um agente virtual



POR VINY MATHIAS

ATUALIZADO: 25 DE MAIO DE 2025 ÀS 16:01

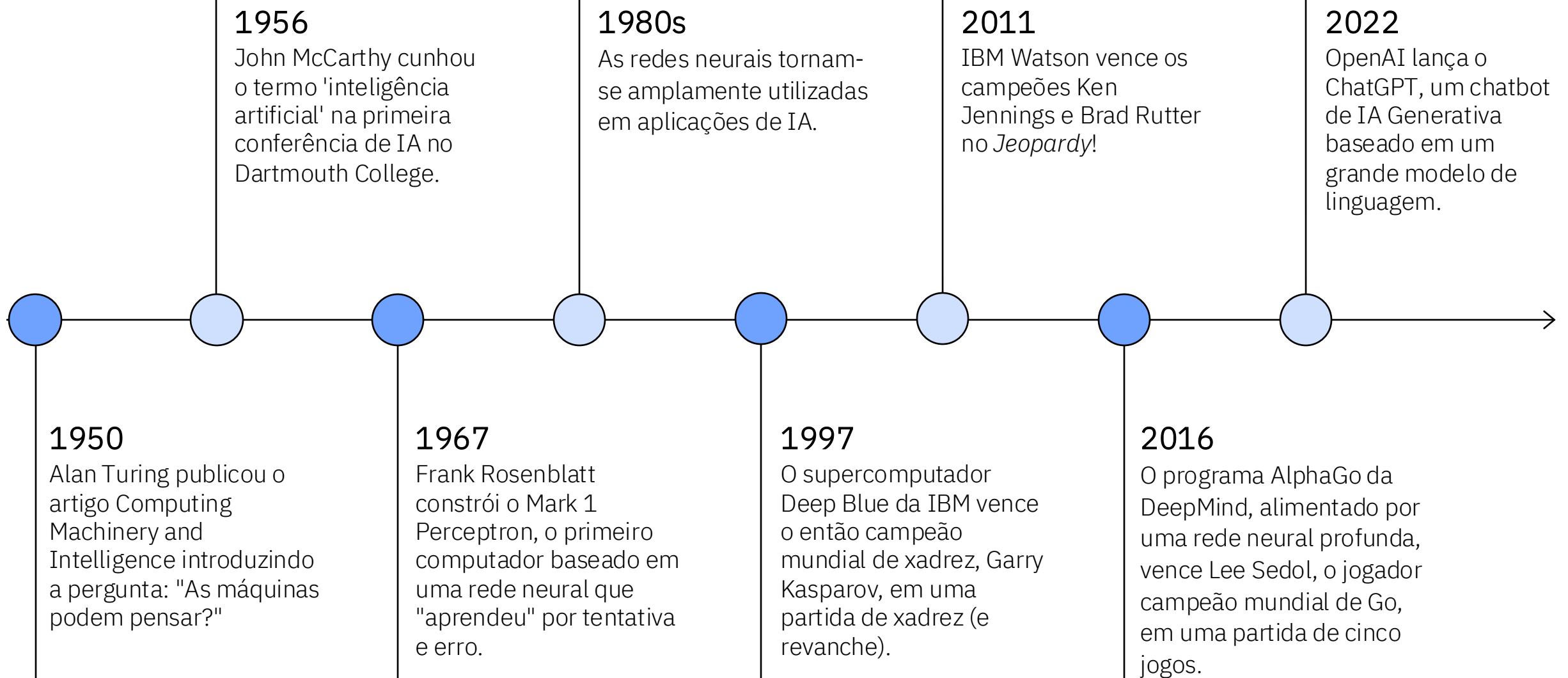


<https://br.ign.com/tech/140588/news/69-dos-membros-da-geracao-z-pedem-coisas-ao-chatgpt-educadamente-com-uma-intencao-oculta>



[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

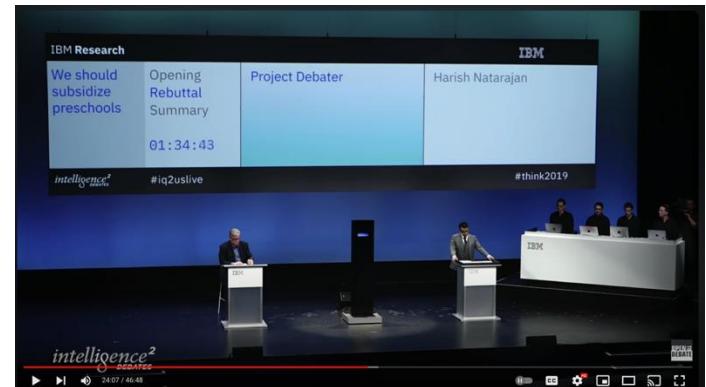
## O Boticário lança 1ºs perfumes feitos com ajuda de inteligência artificial

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

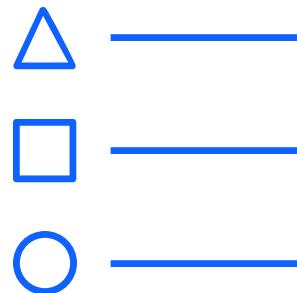


Marco na perfumaria: fragrâncias criadas com ajuda de "robô" chegam dia 27. (Grupo Boticário/Divulgação)

# Project Debater



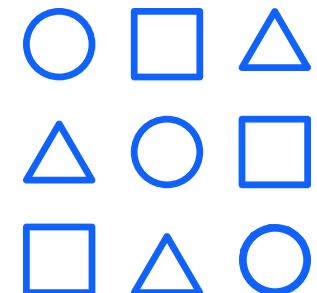
## Sistemas baseados em regras



*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*



A IA pode ser definida como uma técnica que permite que as máquinas imitem funções cognitivas associadas à mente humana – as funções cognitivas incluem todos os aspectos de aprendizagem, raciocínio, percepção e resolução de problemas.

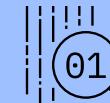
## Machine Learning (ML)

*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).

## Agentes de IA



1950's



1980's



2010's



2020's



2025's



# 82%

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

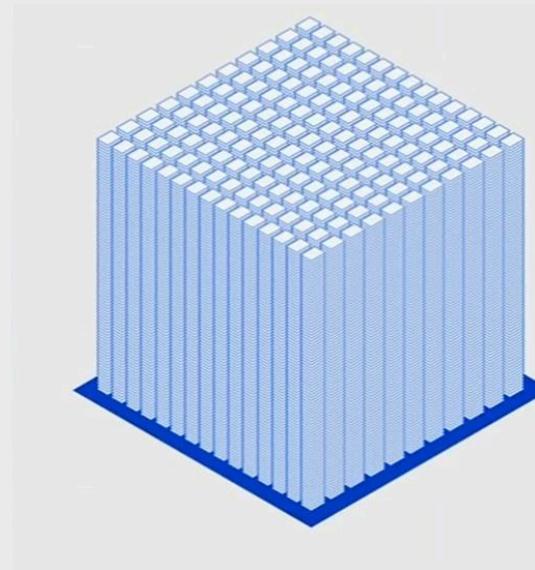
# 55%

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



# 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+

# Employee Experience ↗



Um único canal:  
experiência  
envolvente

2 camadas: Digital &  
Humano



4,7k páginas  
de conteúdo de RH



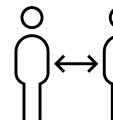
243k  
Usuários  
únicos



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%

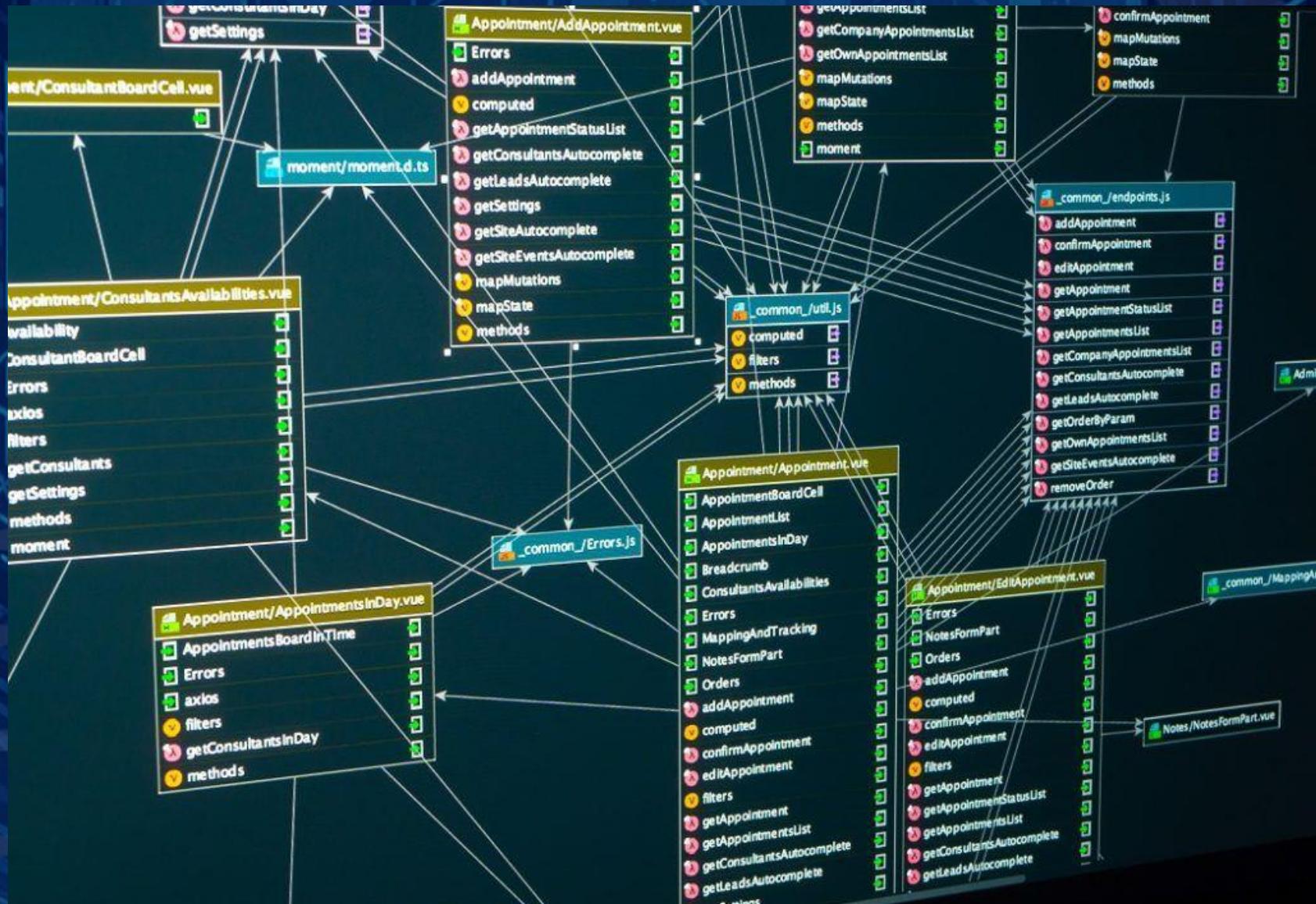
Redução de  
abertura de  
chamados

75%

Feedbacks  
positivos  
(CSAT)

89%

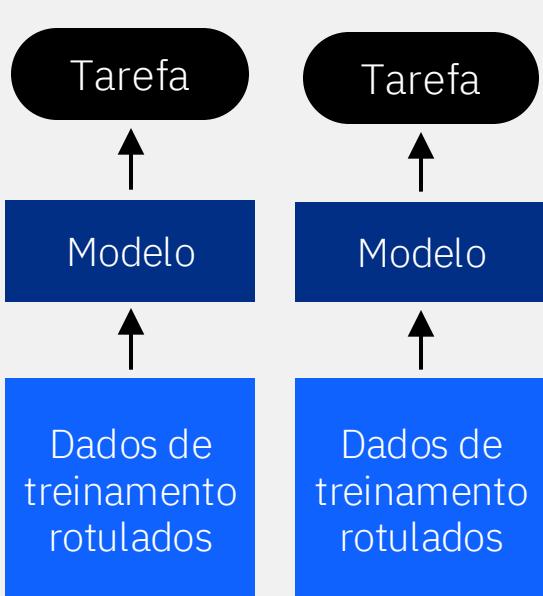
de interações  
resolvidas  
diretamente via  
AskHR



## O 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



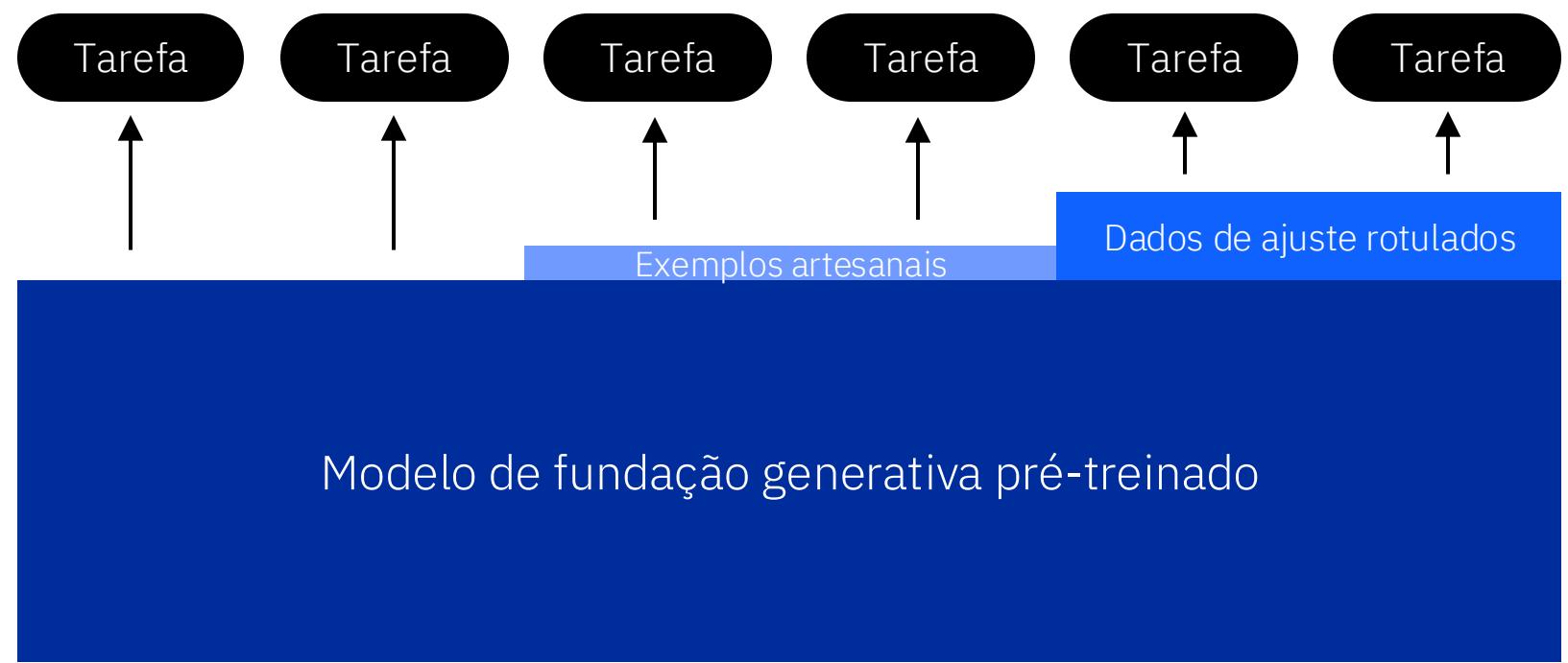
Prompt-tuning

100s – 1000s labeled data points / task

Tarefa

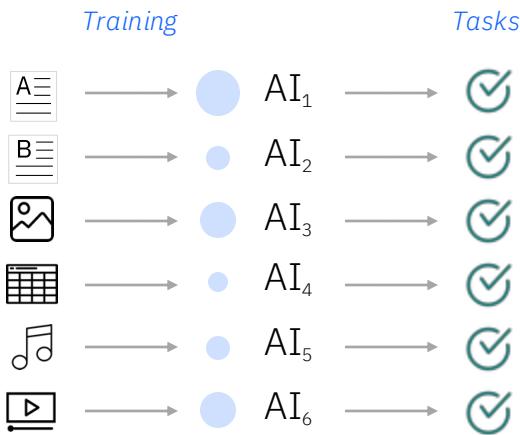


Tarefa



# Foundation models establish a new paradigm for AI capabilities

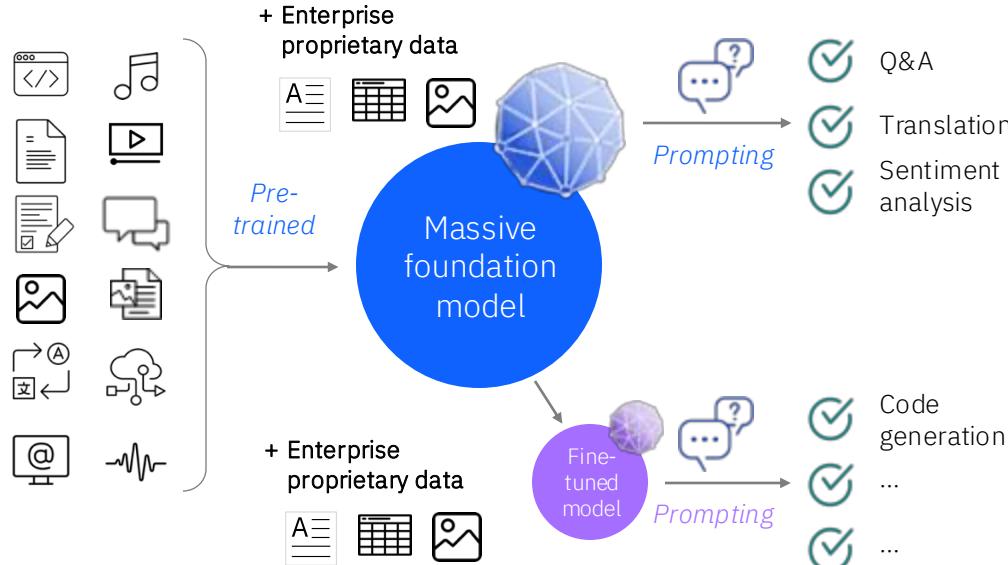
## Traditional AI models



- 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

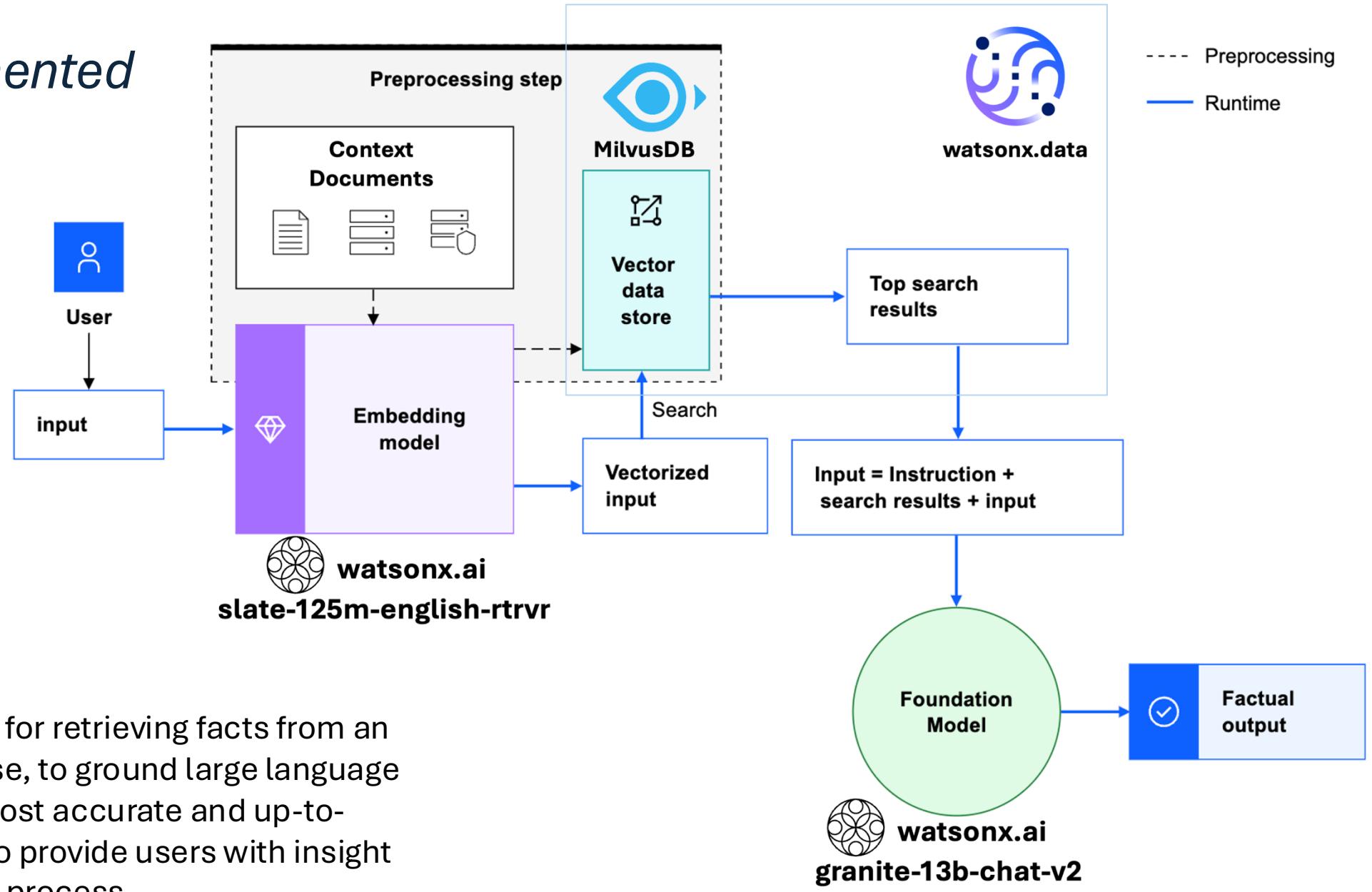
## Key advantages

- 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

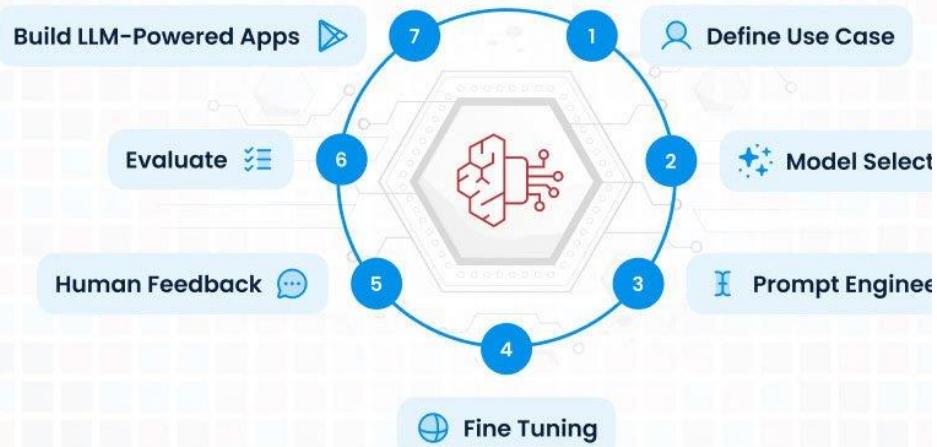
# RAG

## Retrieval Augmented Generation

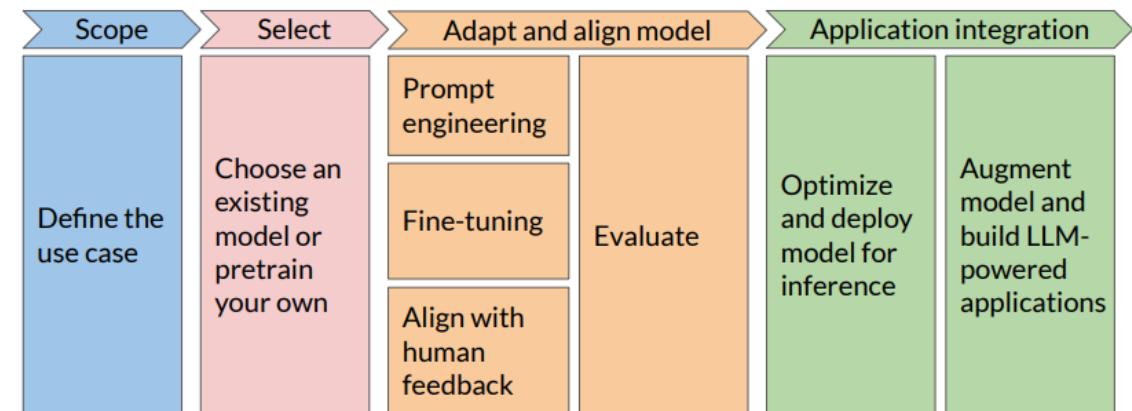


RAG is an AI framework for retrieving facts from an external knowledge base, to ground large language models (LLMs) in the most accurate and up-to-date information, and to provide users with insight into the LLM generating process.

## 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

Israel

Hybrid cloud

Almaden

Tokyo

Security

Brazil

United Kingdom

Cambridge

Yorktown Heights

India

Zurich



6 Nobel  
Laureates



5 National  
Medals of Science



10 Medals of  
Technology



6 Turing  
Awards



A diagram illustrating the relationship between classical and quantum computing components. It consists of three black circles arranged horizontally. The first circle contains the word "Bits". To its right is a plus sign ("+"). The second circle contains the word "Neurons". To its right is another plus sign ("+"). The third circle contains the word "Qubits".

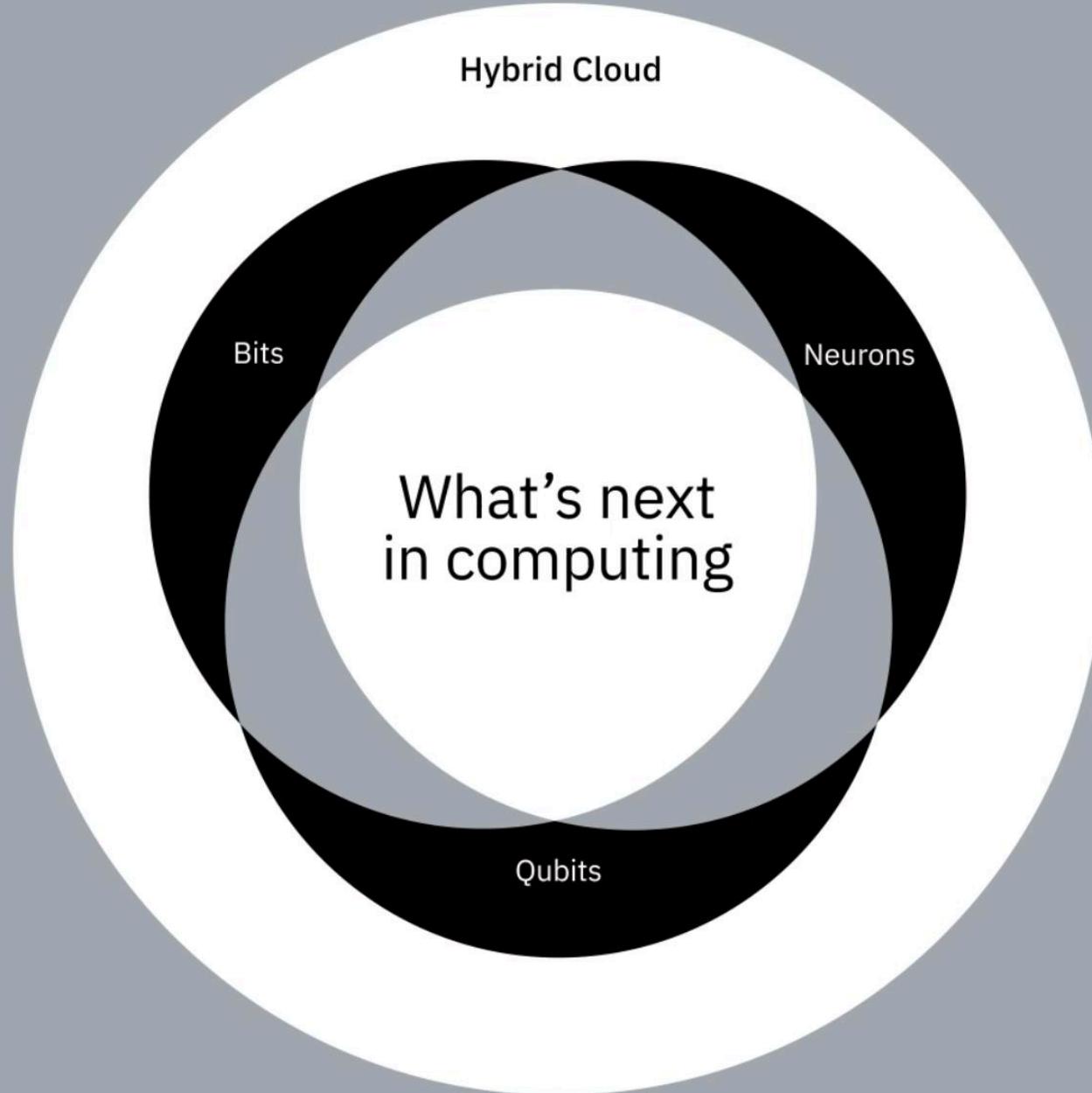
Bits

+

Neurons

+

Qubits



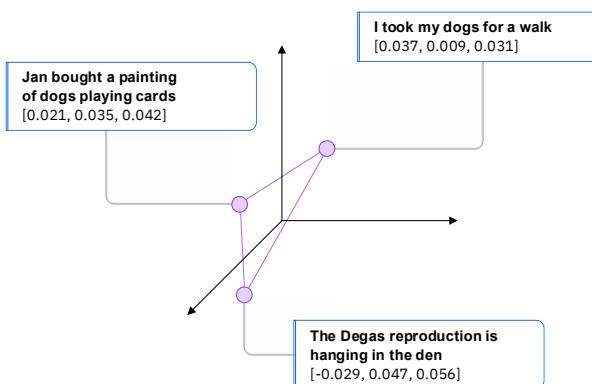
# watsonx AI Platform Engineering

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

## Guardrails framework



## Embedding



## Performance na inferência



\* OpenSource

## Compiler/Driver



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

Rodando e otimizando para AIU/Spyre

# 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

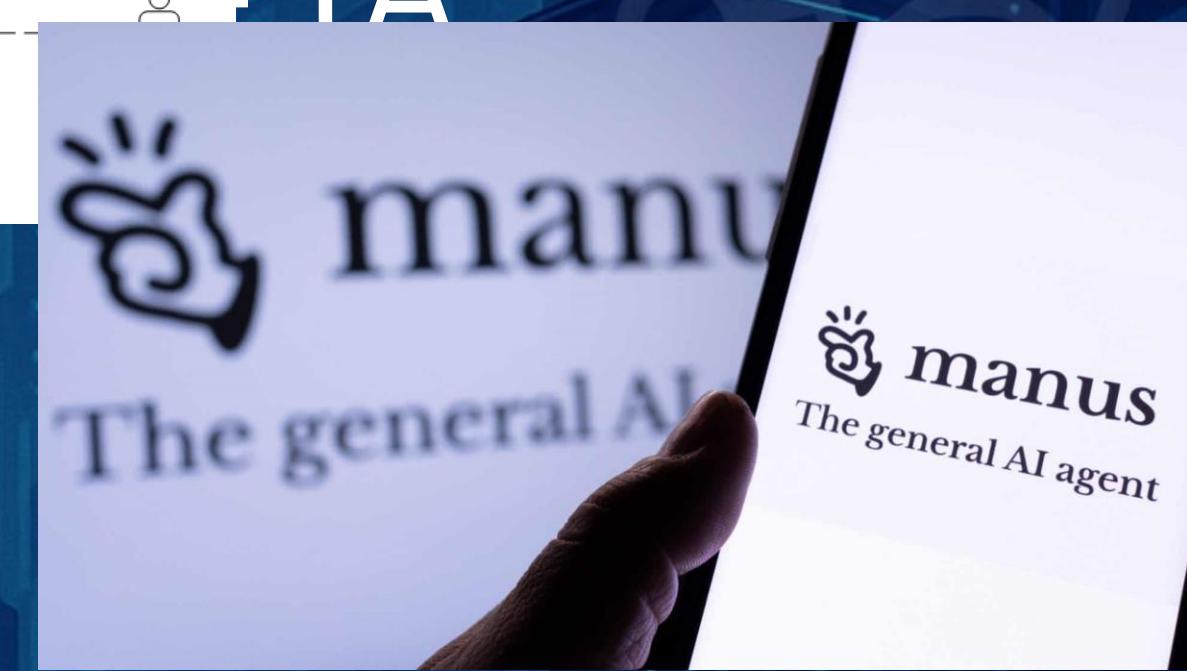
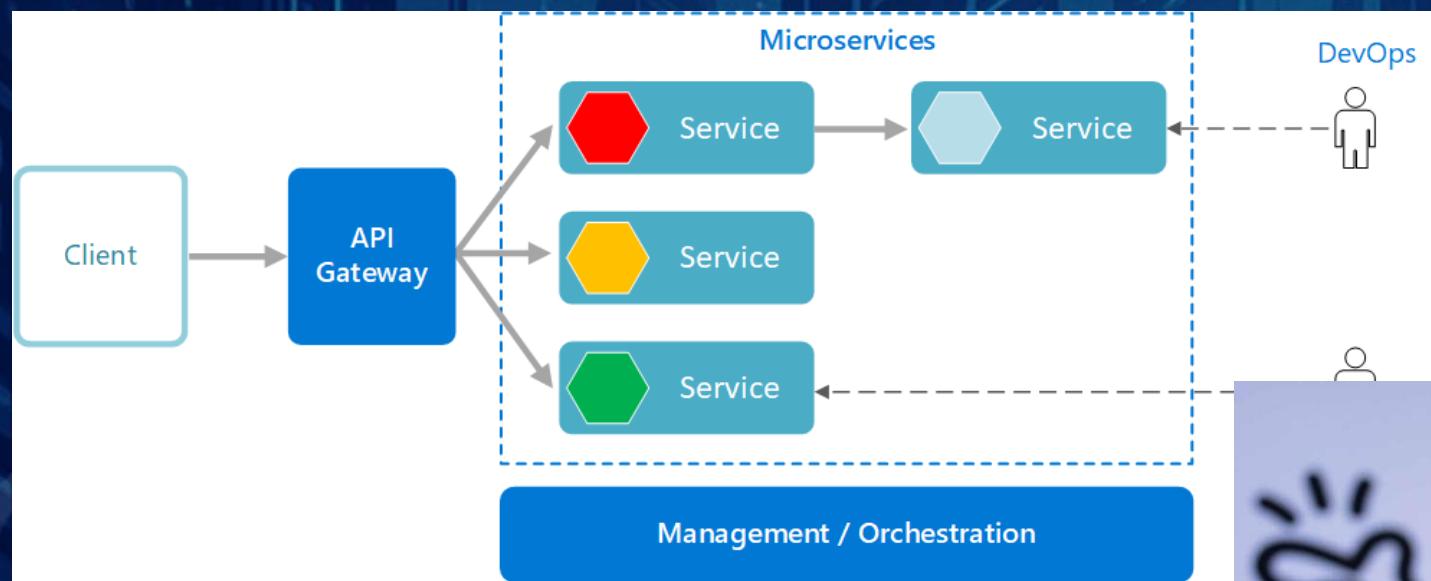
ibm-research models updated by SagiPolaczek 5 months ago

- 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

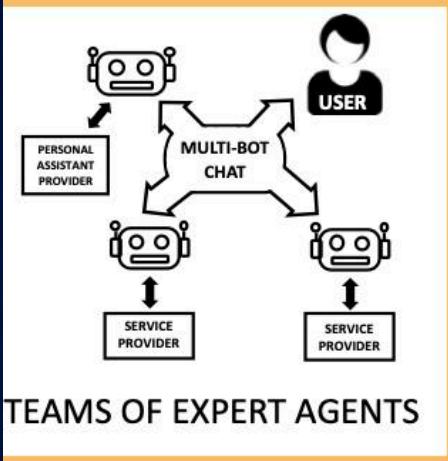
All Models Datasets Spaces

- 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



# As 10 principais tendências tecnológicas estratégicas para 2025





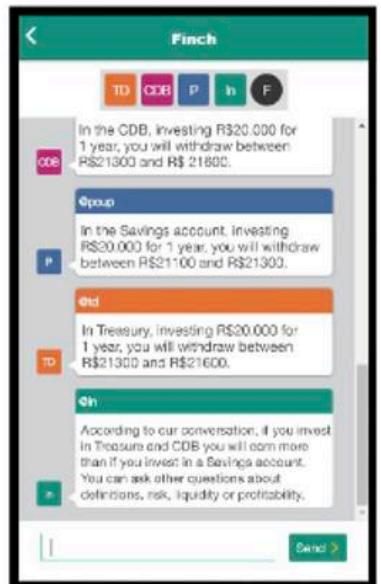
**cdbGuru**  
reactive bot



**poupancaGuru**  
reactive bot



**tesouroGuru**  
reactive bot



User

**investmentGuru**  
proactive bot



from Q&A dialog to multiparty chat

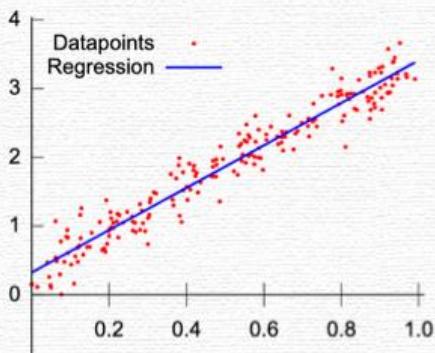
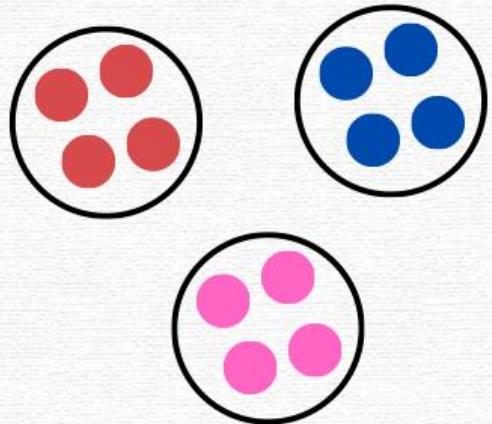
15

Self-service finance advising  
using multiple chatbots

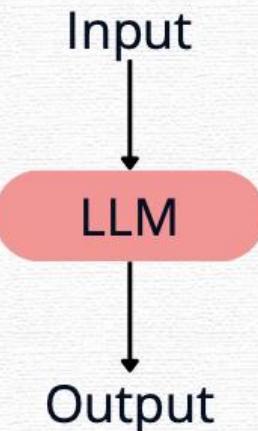
Should we design multi or single bots?  
[DeepDial'18]

# Evolução da Inteligência Artificial

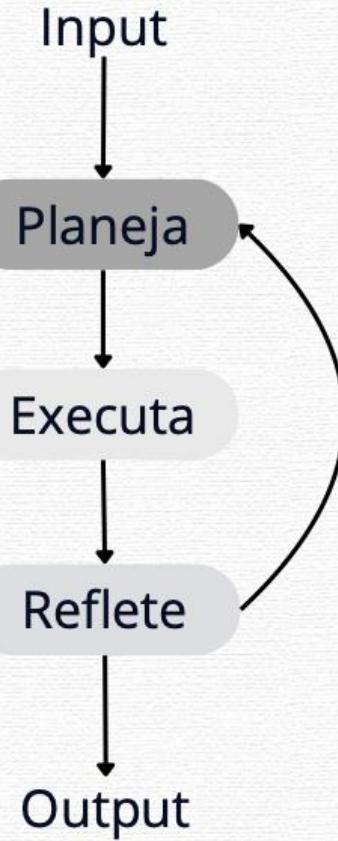
➤ Classificação/Predição



➤ Assistants



➤ AI Agent

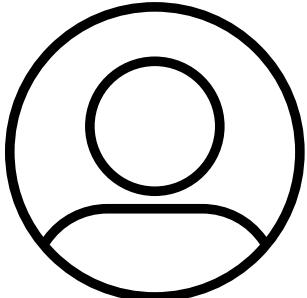


# Introdução

## 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

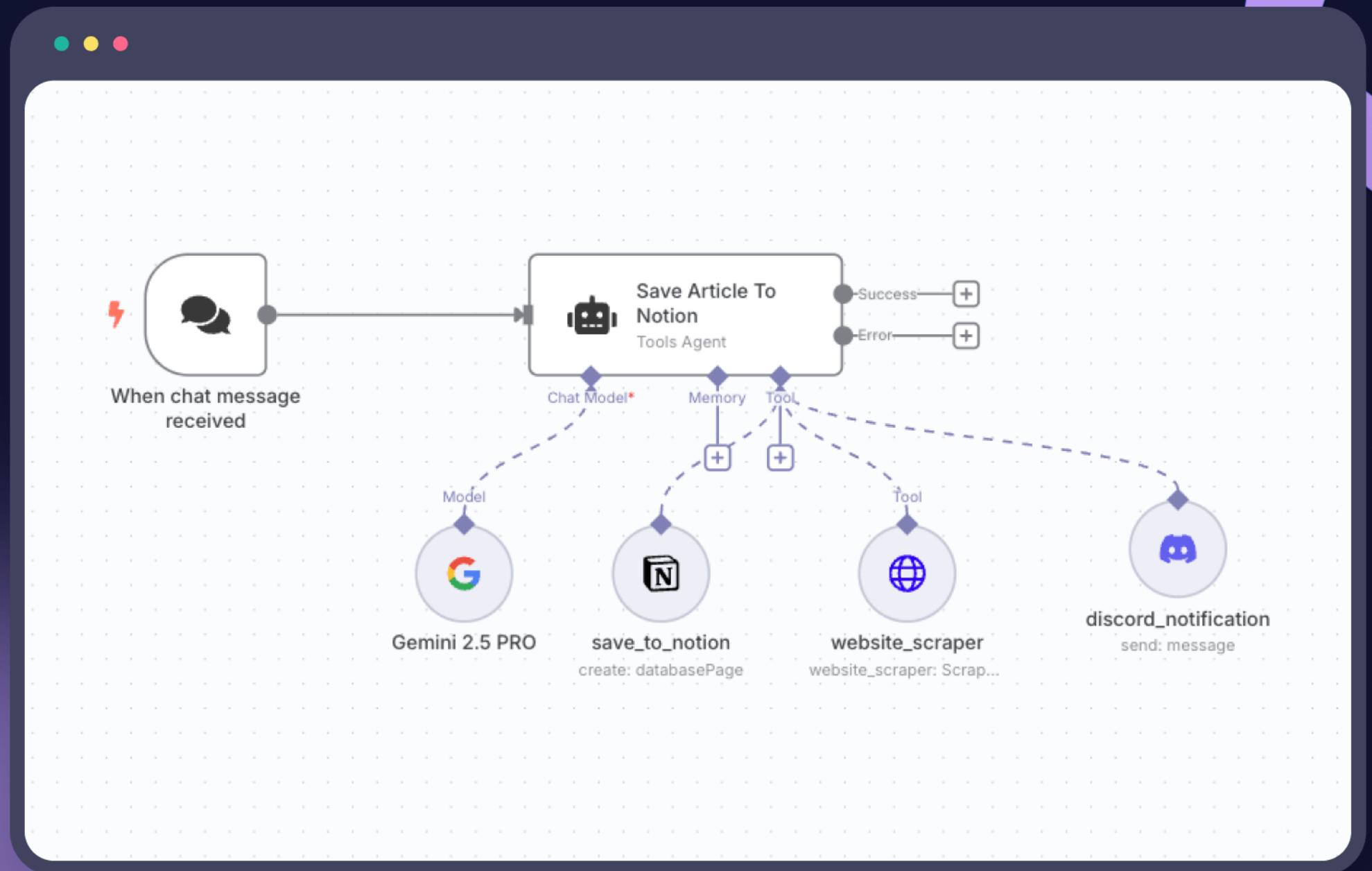
Agent: (tool\_caption) Getting current weather for Las Vegas

Agent: (tool\_input) {"location": {"name": "Las Vegas", "language": "English"}, "start\_date": "2024-09-13", "end\_date": "2024-09-13", "temperature\_unit": "celsius"}  
> Agent: (tool\_output) {"latitude":36.16438,"longitude":-115.14392,"generationtime\_ms":0.08392333984375,"utc\_offset\_seconds":0,"timezone":"UTC","timezone\_abbreviations":["seconds"],"temperature\_2m":22.3,"rain":0,"apparent\_temperature":22.3,"current":{"time":"2024-09-13T11:15","interval":900,"temperature\_2m":22.3,"rain":0,"apparent\_temperature":22.3,"relative\_humidity\_2m":90,"apparent\_temperature":22.3,"hourly":[{"time":["2024-09-13T00:00","2024-09-13T01:00","2024-09-13T02:00","2024-09-13T03:00","2024-09-13T04:00","2024-09-13T05:00","2024-09-13T06:00","2024-09-13T07:00","2024-09-13T08:00","2024-09-13T09:00","2024-09-13T10:00","2024-09-13T11:00","2024-09-13T12:00","2024-09-13T13:00","2024-09-13T14:00","2024-09-13T15:00","2024-09-13T16:00","2024-09-13T17:00","2024-09-13T18:00","2024-09-13T19:00","2024-09-13T20:00","2024-09-13T21:00","2024-09-13T22:00","2024-09-13T23:00"]}], "temperature\_2m": [34.2, 33.6, 31.3, 30.3, 29.7, 28.2, 27.2, 26.2, 24.8, 24.2, 23.3, 22.7, 22.1, 21.5, 20.9, 20.3, 19.7, 19.1, 18.5, 17.9, 17.3, 16.7, 16.1, 15.5, 14.9, 14.3, 13.7, 13.1, 12.5, 11.9, 11.3, 10.7, 10.1, 9.5, 8.9, 8.3], "apparent\_temperature": [30.1, 30.2, 27.9, 26.4, 25.7, 25.3, 23.2, 22.3, 20.7, 19.8, 19.7, 19.1, 18.5, 17.9, 17.3, 16.7, 16.1, 15.5, 14.9, 14.3, 13.7, 13.1, 12.5, 11.9, 11.3, 10.7, 10.1, 9.5, 8.9, 8.3], "units":{"time":"iso8601","apparent\_temperature\_max":27.9,"apparent\_temperature\_min":16.1,"sunrise":"iso8601","sunset":"iso8601"}, "daily":{"time":["2024-09-13T00:00","2024-09-13T01:00","2024-09-13T02:00","2024-09-13T03:00","2024-09-13T04:00","2024-09-13T05:00","2024-09-13T06:00","2024-09-13T07:00","2024-09-13T08:00","2024-09-13T09:00","2024-09-13T10:00","2024-09-13T11:00","2024-09-13T12:00","2024-09-13T13:00","2024-09-13T14:00","2024-09-13T15:00","2024-09-13T16:00","2024-09-13T17:00","2024-09-13T18:00","2024-09-13T19:00","2024-09-13T20:00","2024-09-13T21:00","2024-09-13T22:00","2024-09-13T23:00"]}, "ise": ["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.



CEOs must consider five mindshifts to create clarity in crisis

41

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



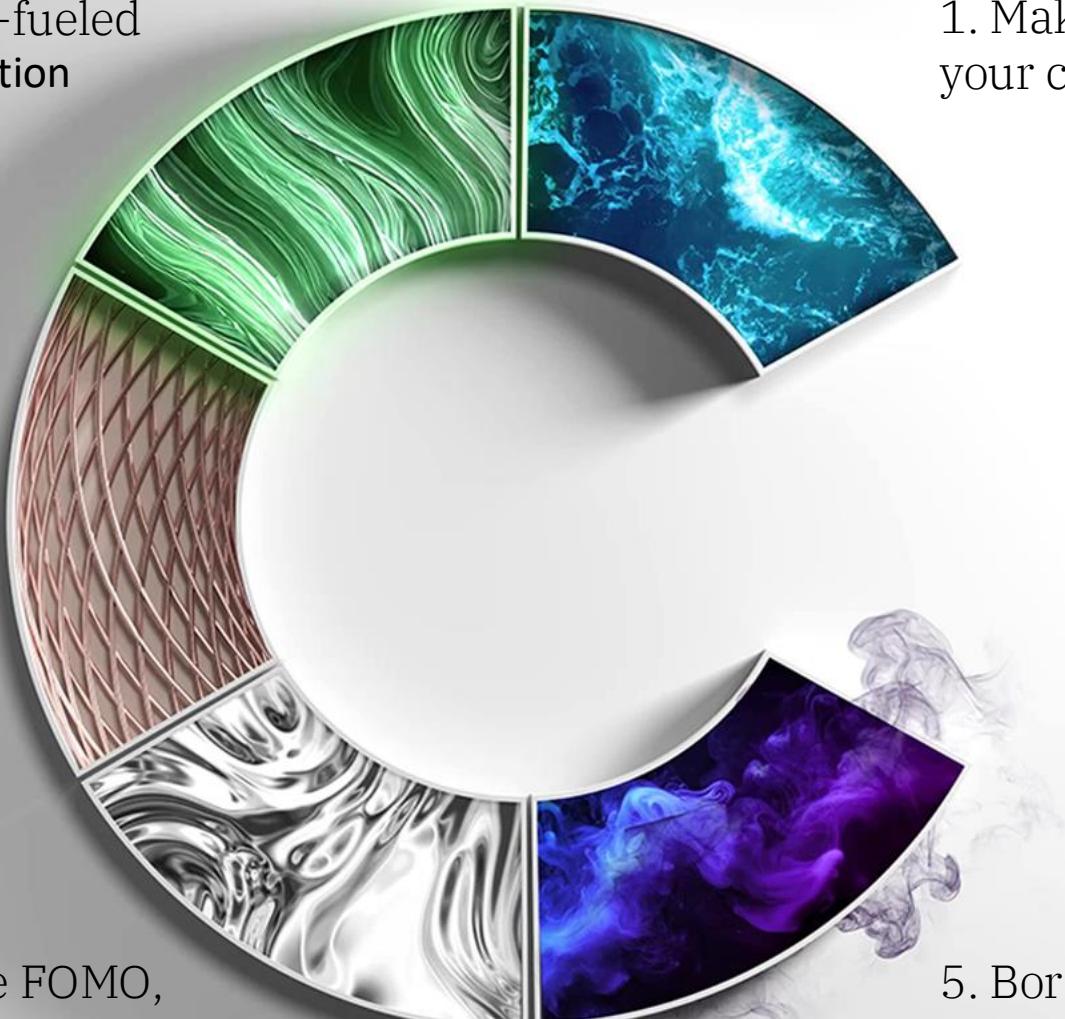
2. Embrace AI-fueled creative destruction

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



“Os pessimistas parecem inteligentes, mas os otimistas ganham dinheiro. Acho que há oportunidades quando há deslocamentos de mercado e, se você sentir que tem uma vantagem de informação, deve se inclinar para o risco.”

[Stephen J. Rich](#)

Chairman, CEO and President  
Mutual of America Financial Group



18TH EDITION

# 2025 TECH TRENDS REPORT

FTSG



[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.



[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)



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

## FUTURE TRENDS



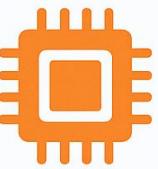
OpenSource  
models



SLMs  
Small Language



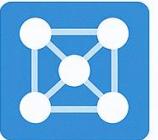
Agentic AI



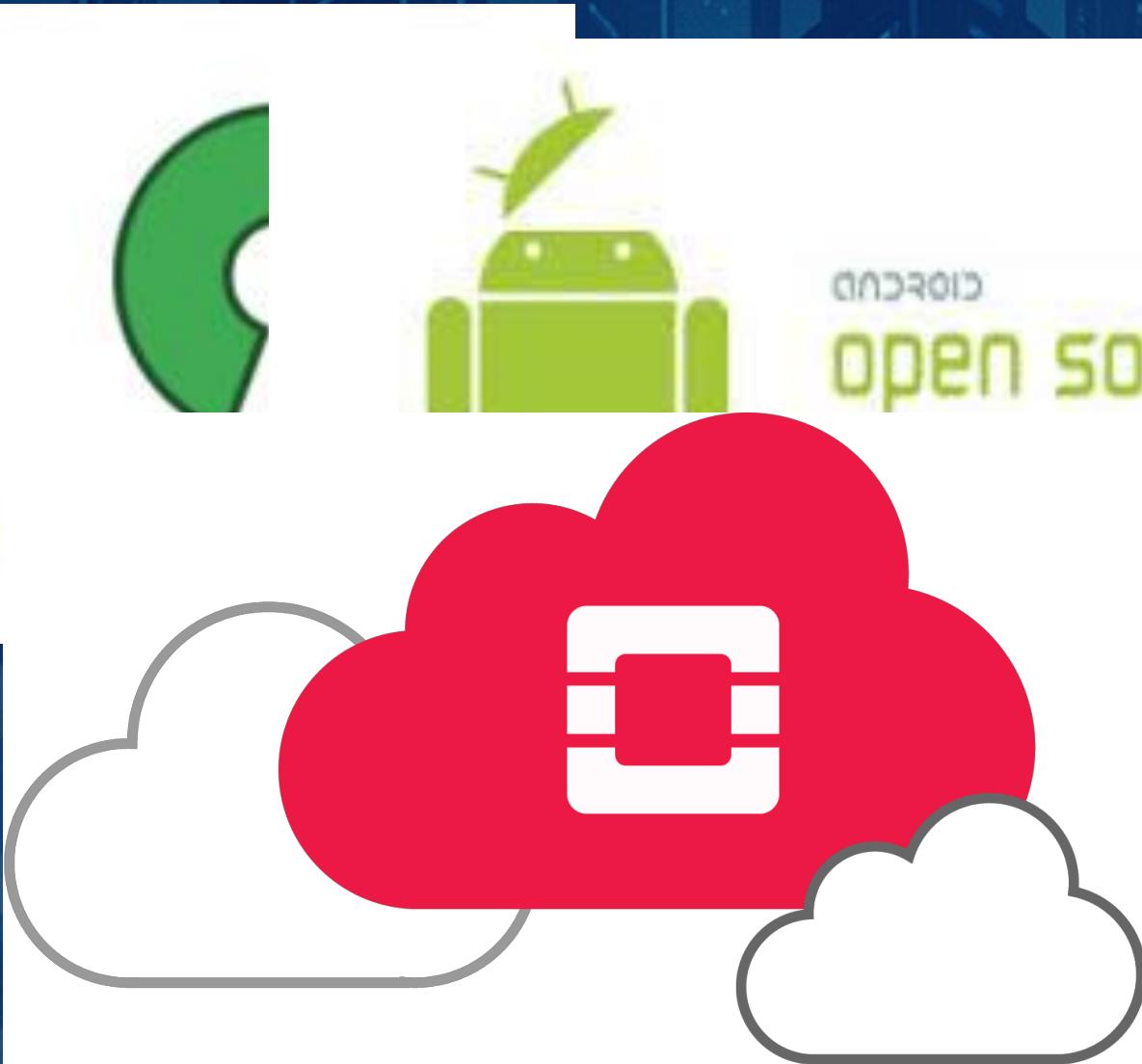
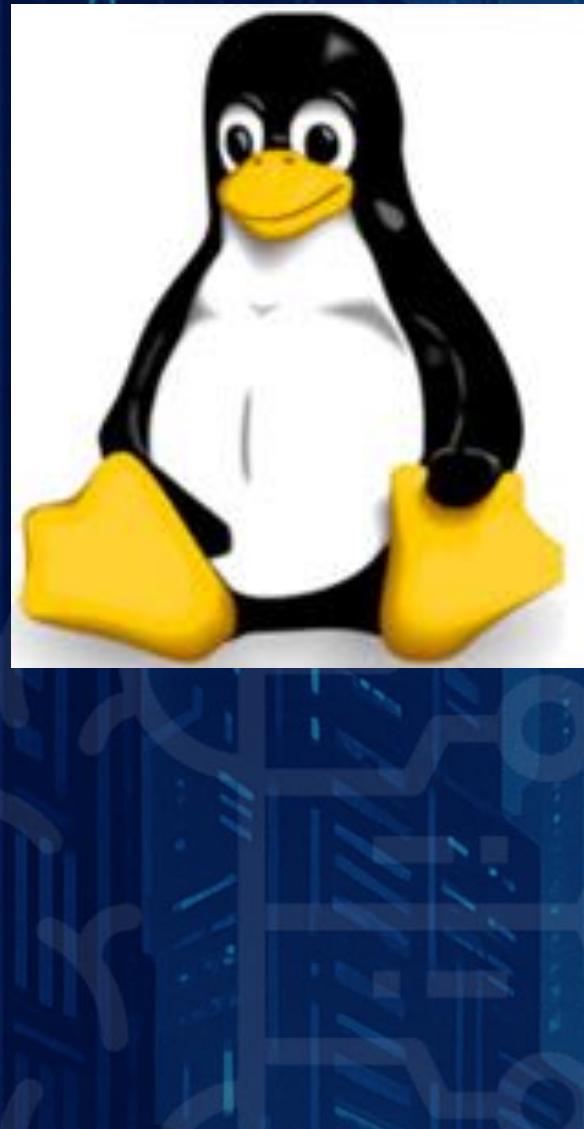
Specific  
hardware



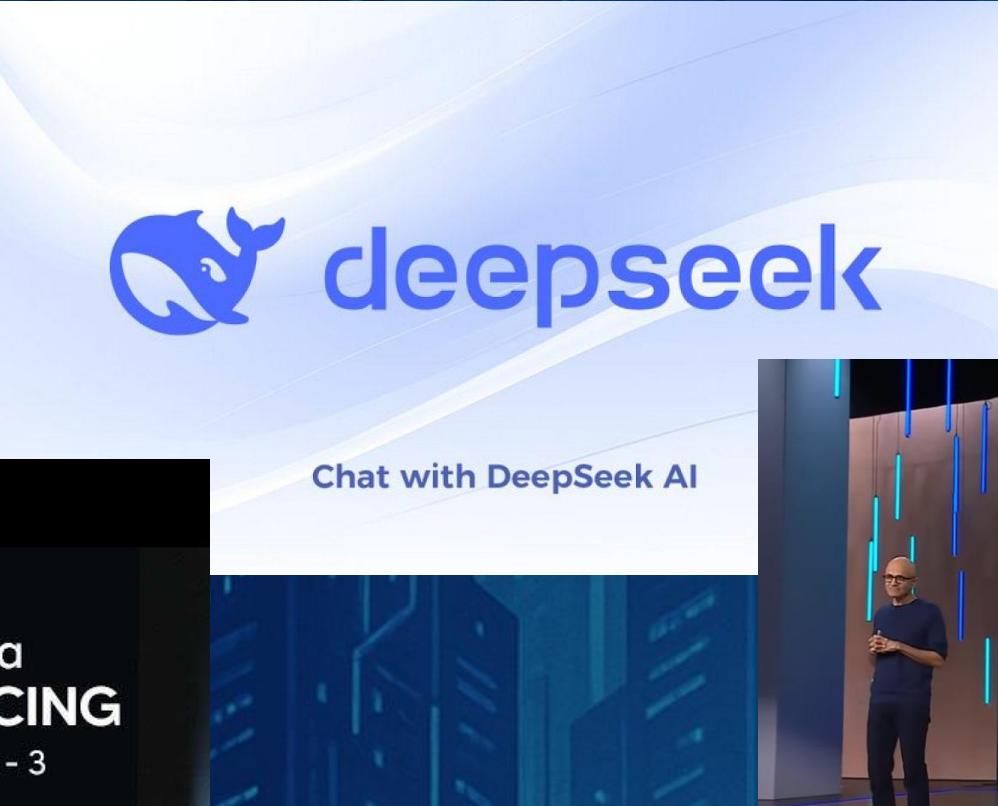
AI Platforms



Quantum  
computing



# OpenSource Models



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

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

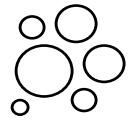
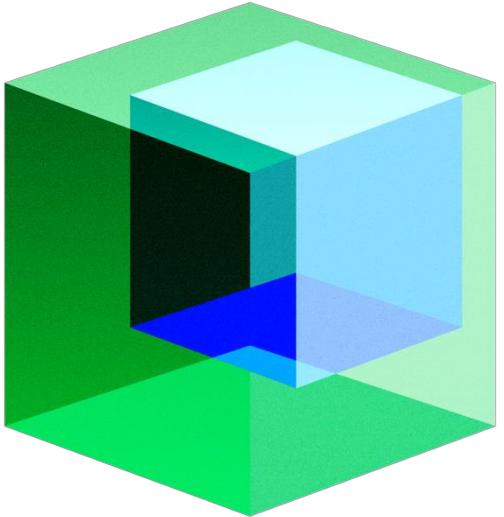
**CALCULAR**

[Não sei meu CEP](#)

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

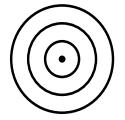
# IBM Granite

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 links. The main area shows a conversation with an AI model named "granite3-moe:1b". A user asks: "Como você explicaria inteligência artificial brevemente para minha avó de 89 anos?". The AI responds: "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." Below the response are several small icons for editing, sharing, and other functions. At the bottom of the screen, there is a footer with links for "Send a Message", "Web Search", "Code Interpreter", and a microphone icon for voice input.

chat.alanbraz.com.br/c/d267d8d8-2979-4cf9-af99-e14a2ea14e9b

New Chat

Notes

Workspace

Search

Chats

Previous 7 days

AI Explained for an 89-Year-Old

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

Send a Message

Alan Braz

Web Search

Code Interpreter

AB

granite3-moe:1b

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

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.



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

alanbraz.com.br / pullrecast.dev

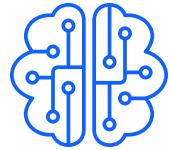


*Agentes de ia / ai agents / agentic ai*



As capacidades da  
IA estão crescendo rapidamente

IA que prevê



*Machine learning*

IA que cria



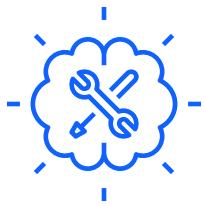
*IA Generativa*

IA que conversa



*Assistentes de IA*

IA que trabalha



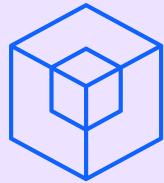
*Agentes de IA*

# A evolução da IA conversacional



## Chatbots

- Baseado em regras
- Determinístico
- Pré-programado



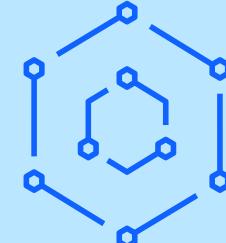
## Assistentes virtuais

- Processamento de Linguagem Natural (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



## AI Agents

Streamline repetitive tasks



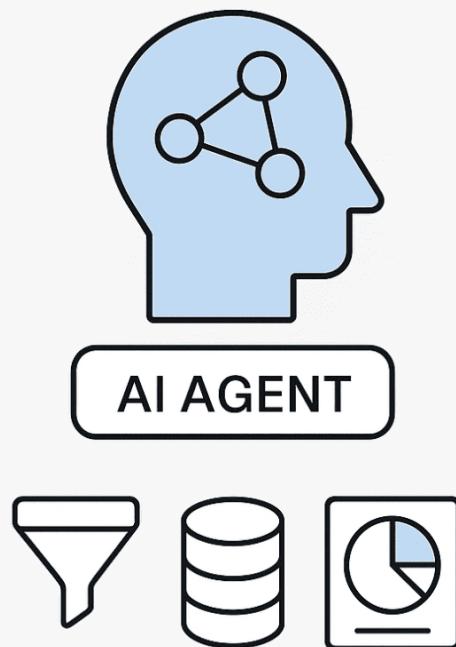
## Agentic AI

Manage complex projects

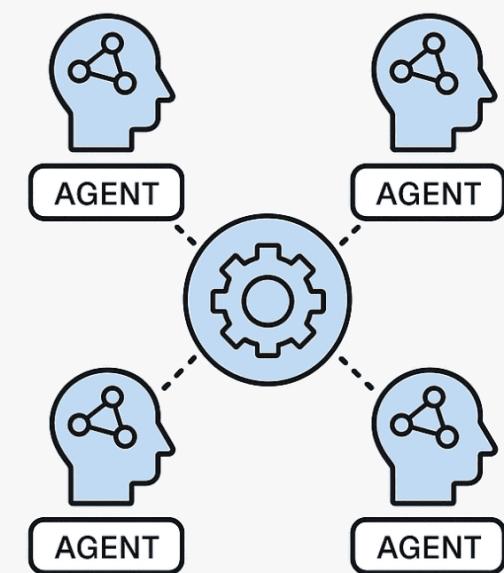
<https://www.mygreatlearning.com/blog/ai-agents-and-agentic-ai/>

<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/>

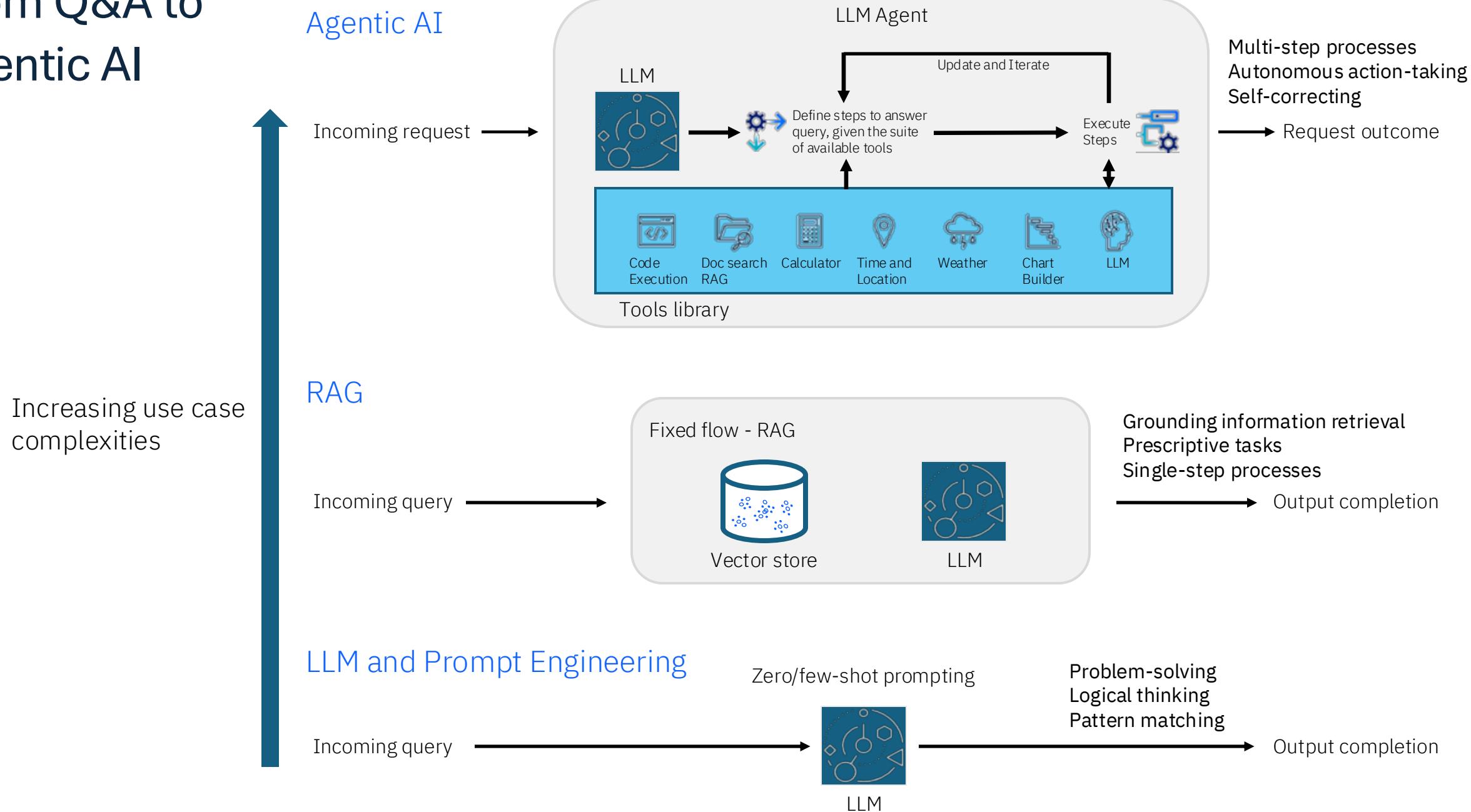
## AI Agents



## Agentic AI

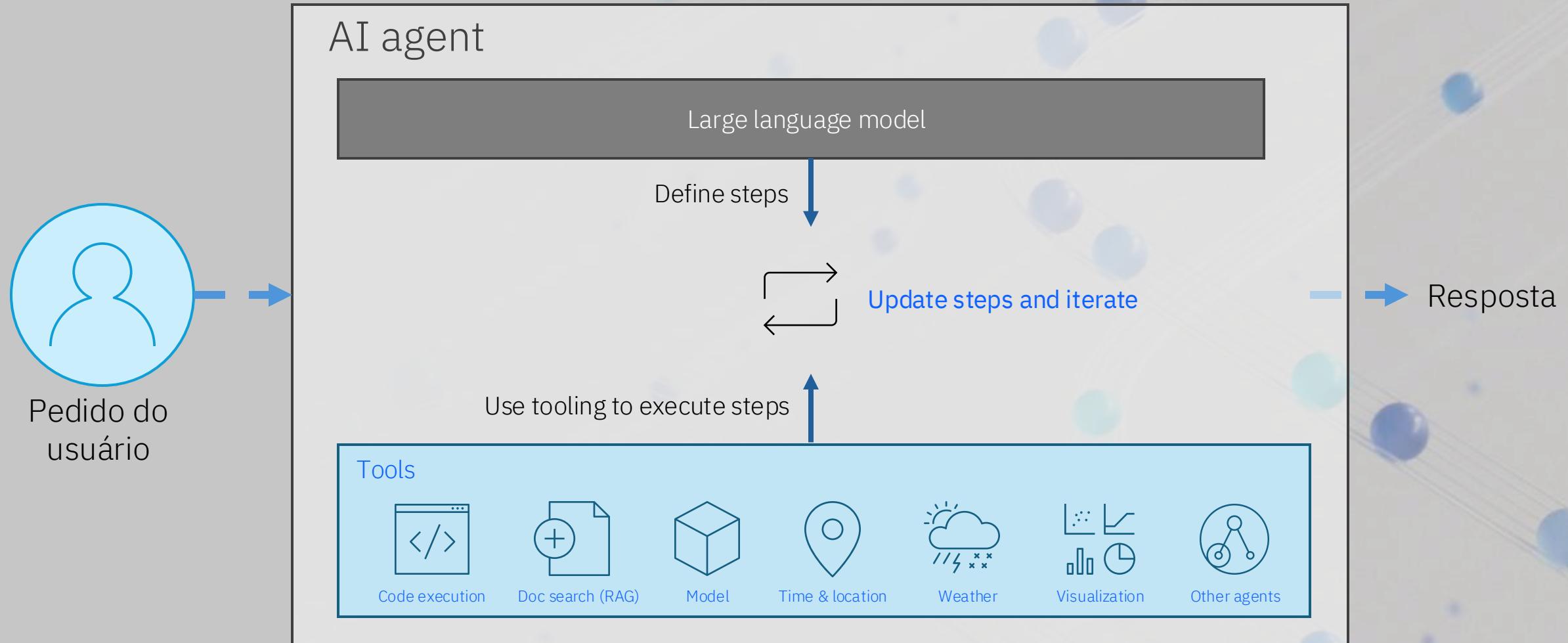


# From Q&A to Agentic AI



# O que são AI Agents e como funcionam?

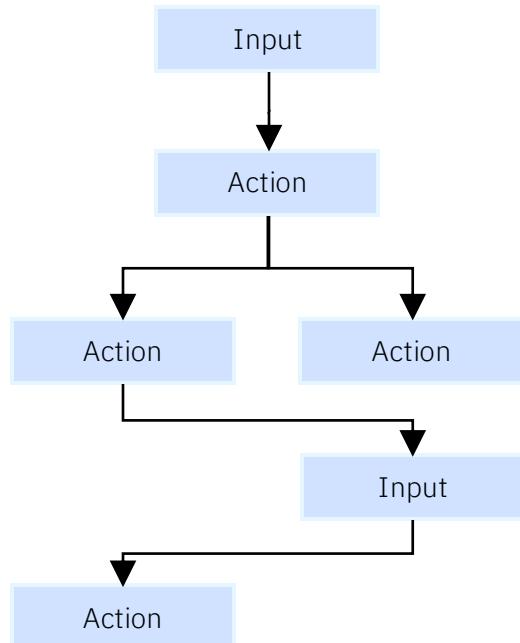
Um AI agent é um programa cuja lógica de execução é controlada por LLM



# Do Assistente ao Multi-Agent

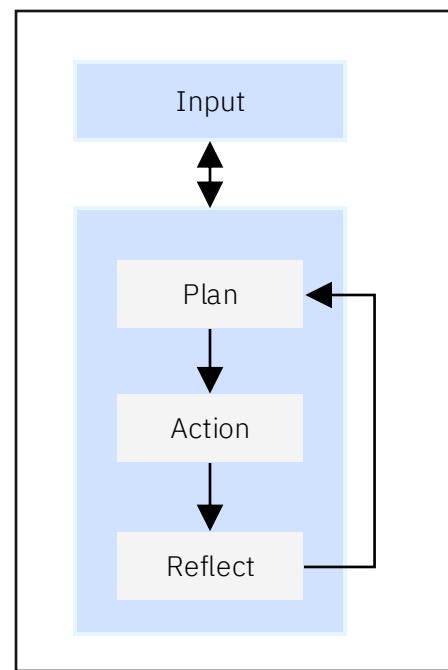
## Assistentes tradicionais

- Baseado em regras (se x, faça y)
  - Caminhos de ação predefinidos



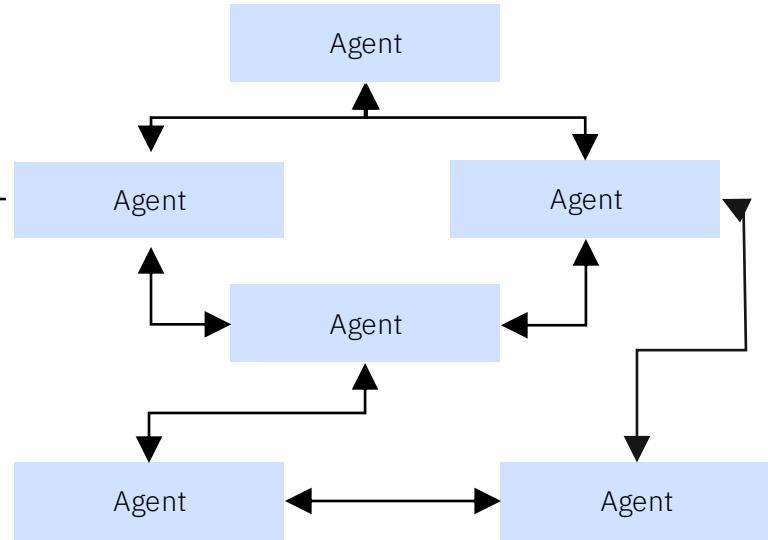
## Assistentes Single-agent

- Baseados em tarefas (ex.: reserva de voo)
  - Restrições de desempenho conforme aumenta o escopo da tarefa
  - Controle limitado

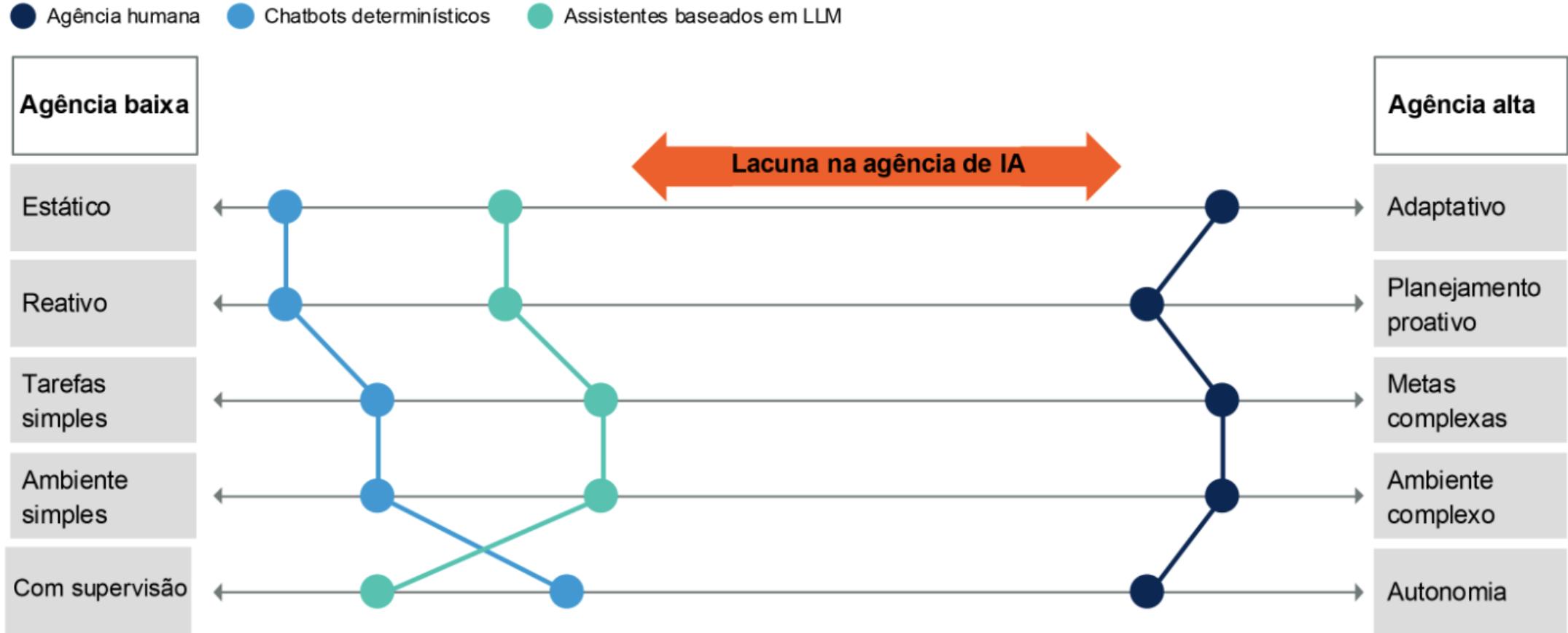


# Assistentes Multi-agent

- Baseados em domínio (ex.: agente de viagens)
  - Agentes especializados (planejador, reservador de voos, reservador de hotéis, etc.) trabalham juntos, melhorando o desempenho do sistema
  - Controle sobre como os agentes se comunicam
  - Múltiplas opções de arquitetura (em rede, hierárquica, supervisor, personalizada)



# A lacuna na adoção da Agentic AI é preocupante



Fonte: Gartner

806843\_C

# Capacidades da IA estão crescendo rapidamente

IA que cria



*Generative AI*

IA que prediz



*Machine learning*

IA que conversa



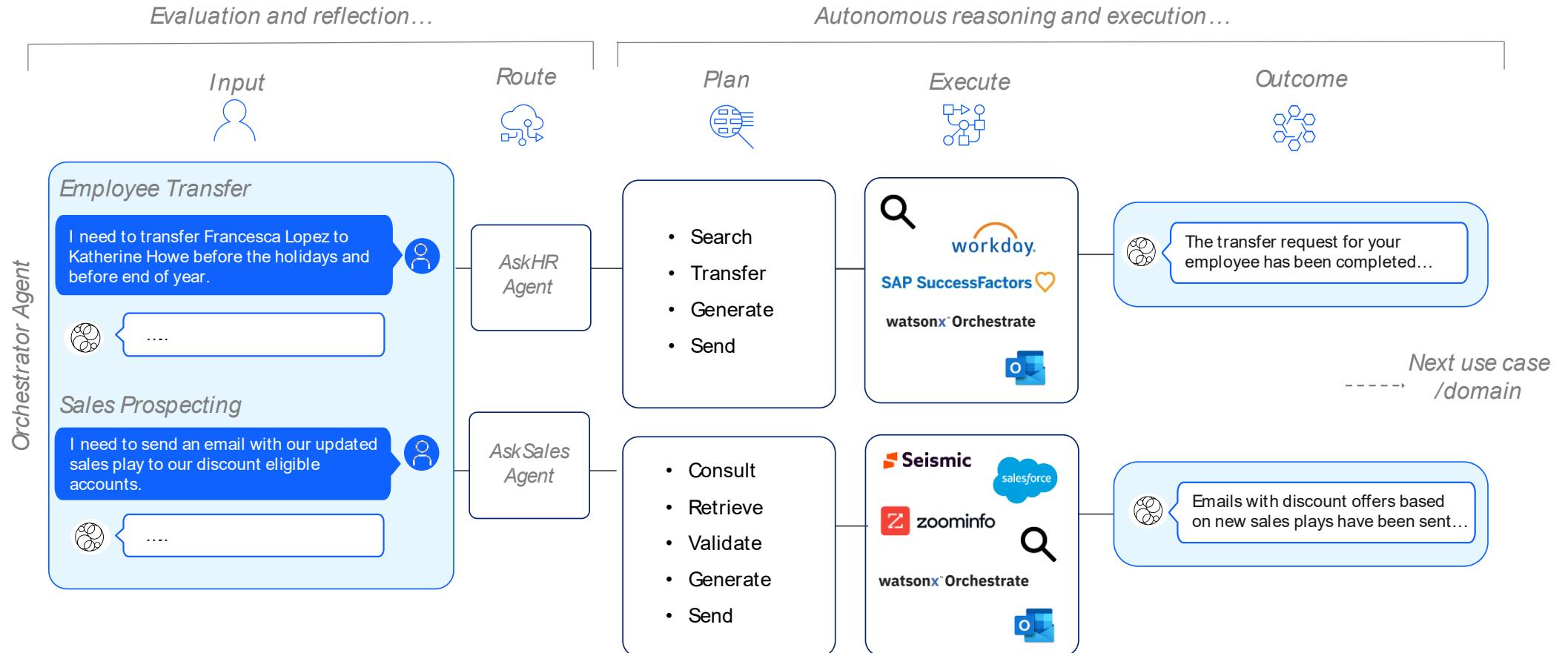
*AI assistants*

IA que trabalha

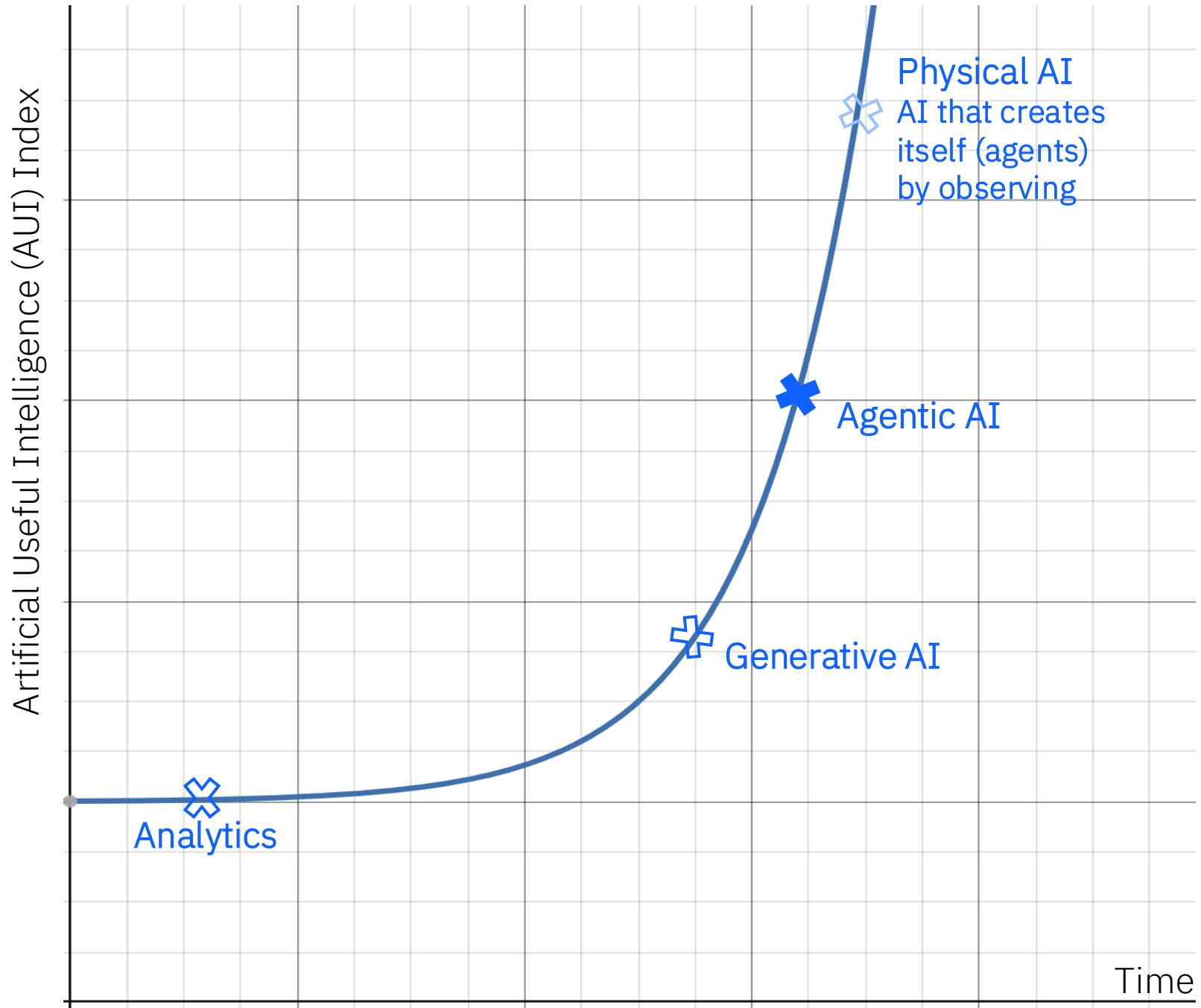


*AI agents*

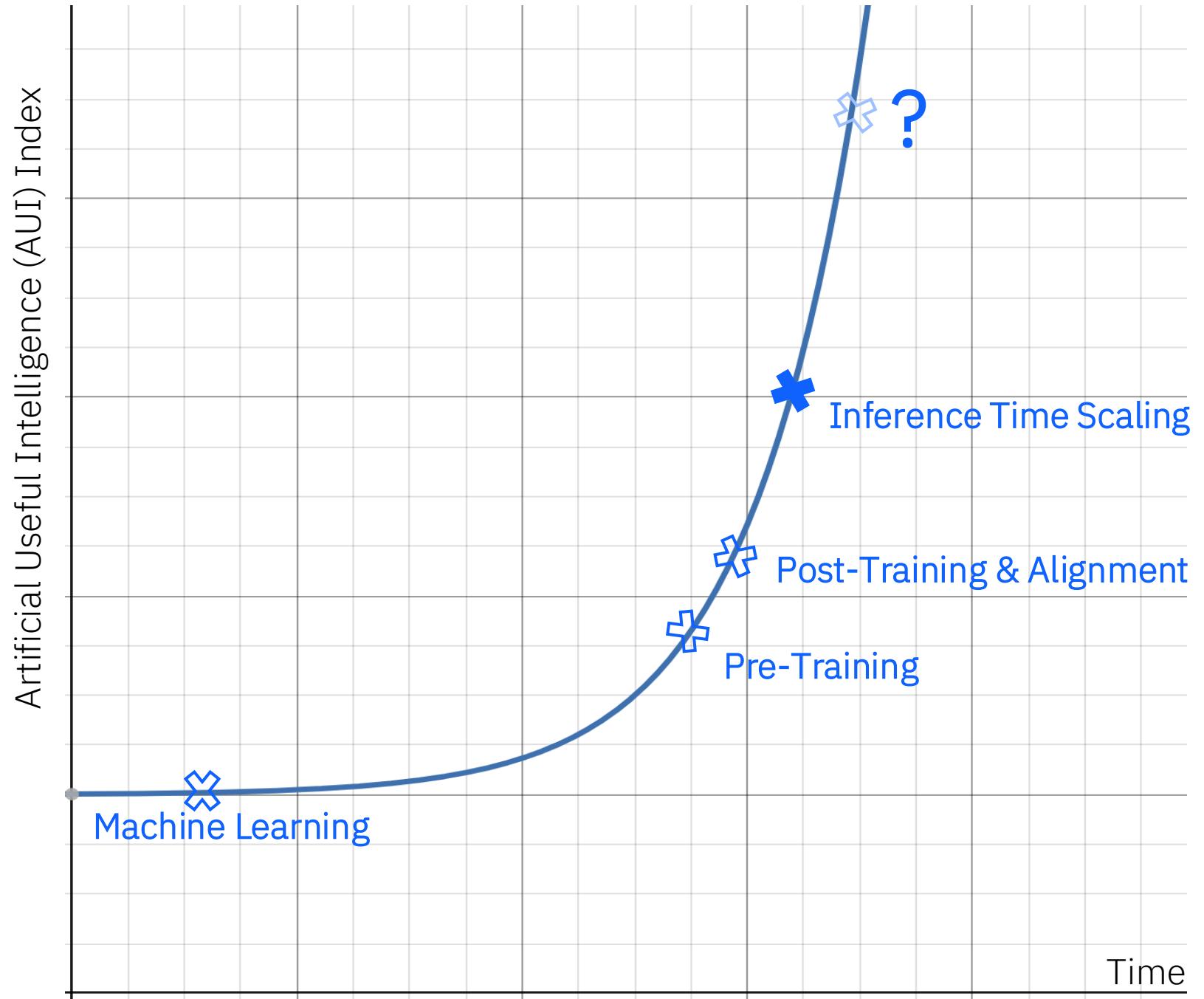
# Orchestrator Agent



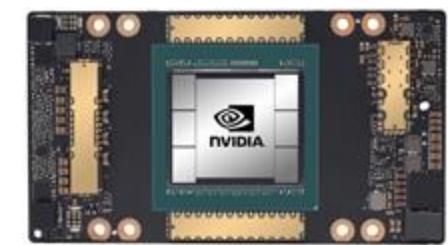
# Artificial Intelligence Futures



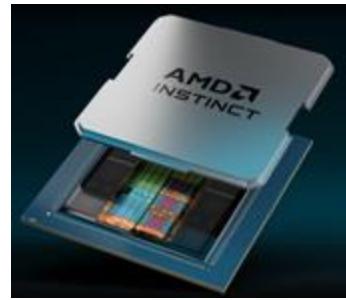
# AI Technology Futures



# Hardware Support



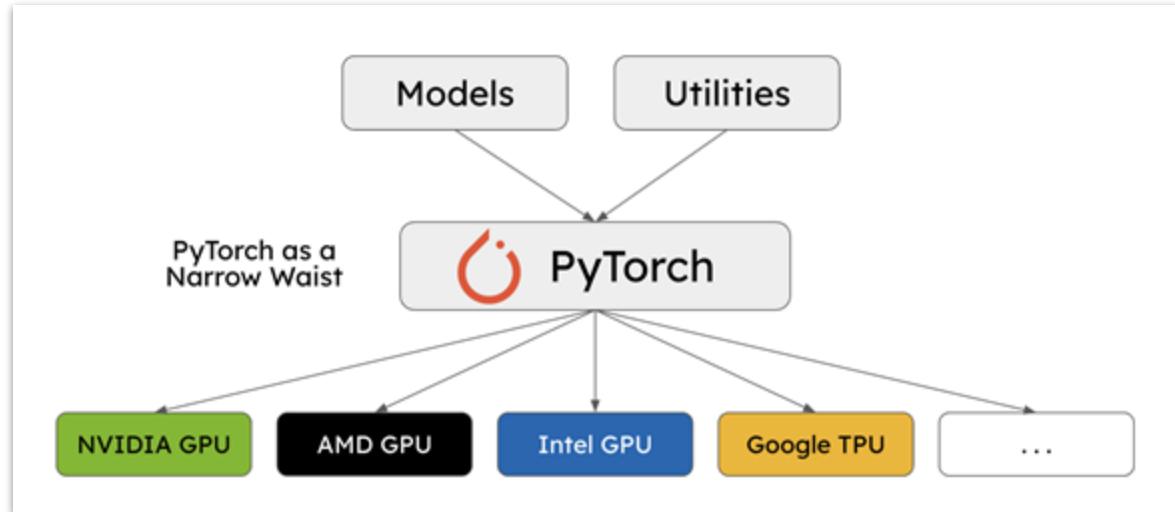
NVIDIA GPU



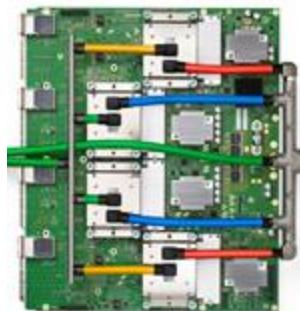
AMD GPU



AWS Neuron



Intel CPU/GPU



Google TPU



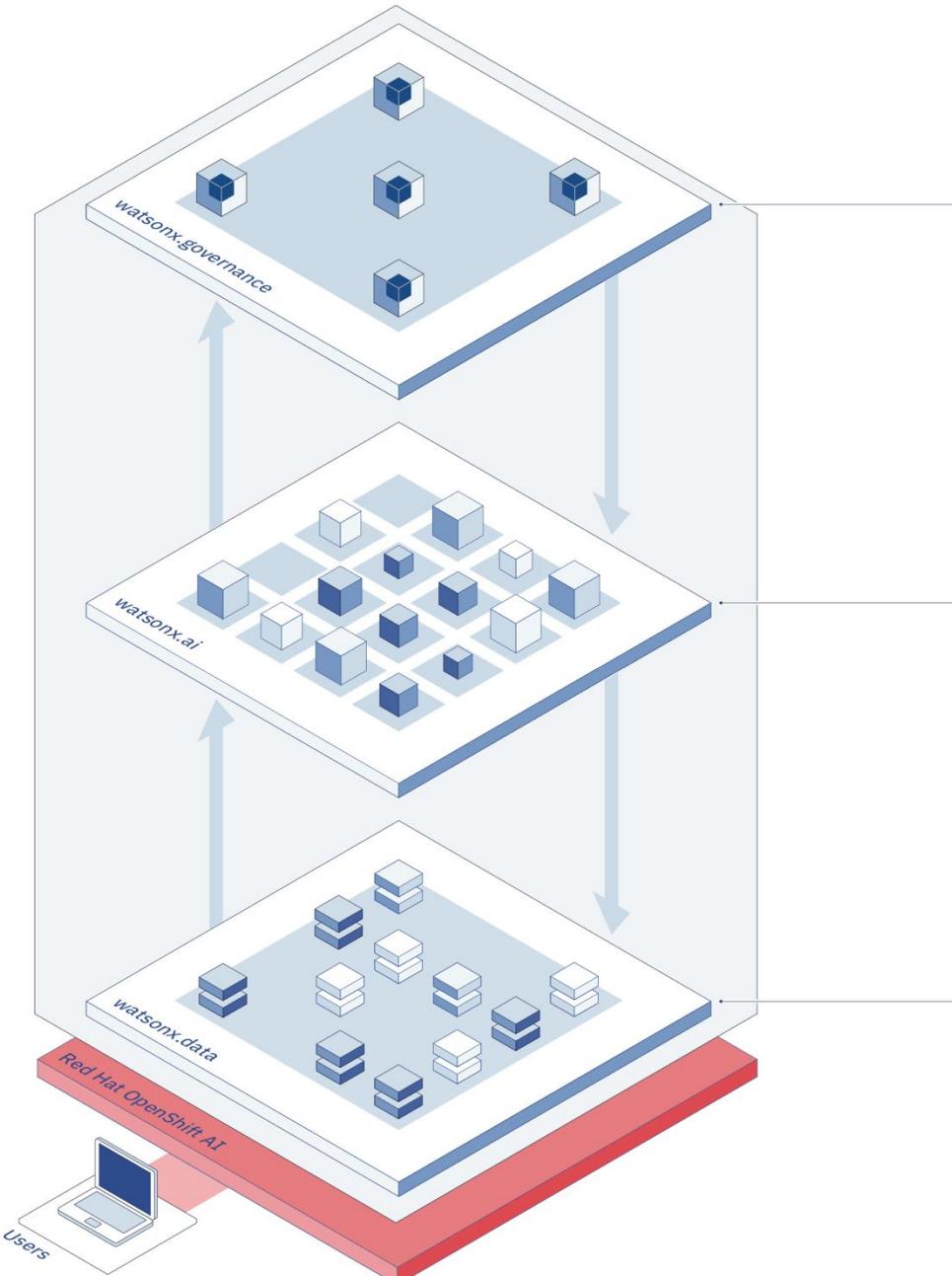
Intel Gaudi



IBM Spyre

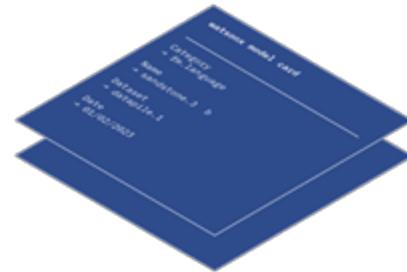


# watsonx, the platform for AI value creators



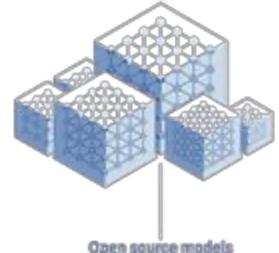
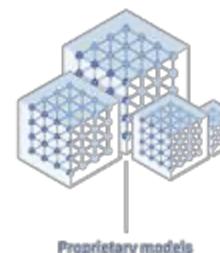
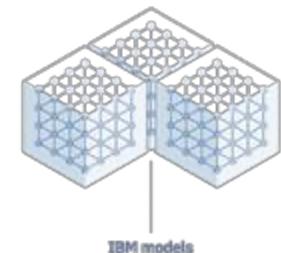
## watsonx.governance

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



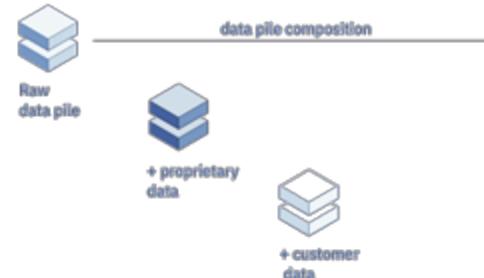
## watsonx.ai

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

The future of computing is quantum.

# 2024

**Expand the utility of quantum computing.**

Demonstrate accurate execution of a quantum circuit at a scale beyond exact classical simulation. (5K gates on 156 qubits)

▼ Learn more

# 2025

**Extend algorithms on quantum computing + HPC and demonstrate error correction code.**

We will release Quantum + HPC tools that will leverage Nighthawk, a new higher-connectivity quantum processor able to execute more complex circuits.

▼ Learn more

# 2026

**Demonstrate first example of scientific quantum advantage and a fault-tolerant module.**

We will demonstrate the first examples of quantum advantage using a quantum computer with HPC.

▼ Learn more

# 2027

**Diversify quantum advantage and entangle fault-tolerant modules.**

The scale, quality, speed of the quantum computer will improve to allow executing quantum circuits at a scale of 10K gates on a 1000+ qubits.

▼ Learn more

# Mais de 40% dos projetos de IA Agêntica serão cancelados até o final de 2027

devido ao aumento dos custos, valor comercial pouco claro ou controles de risco inadequados, de acordo com o Gartner, Inc.

<https://www.linkedin.com/pulse/40-dos-projetos-de-agentic-ai-ser%C3%A3o-cancelados-gilberto-ouoff/>

Fonte: [Gartner](#)

A maioria dos projetos de IA Agêntica no momento são **experimentos** em estágio inicial ou **provas de conceitos** que são impulsionados principalmente por **hype** e muitas vezes são mal aplicados

pelo menos 15% das decisões de trabalho do dia-a-dia serão tomadas de forma autônoma até 2028

33% dos aplicativos de software corporativo incluirão IA agêntica até 2028

# E agora?!?!

Estamos prontos  
para esse  
Futuro?



[mentoría@pullrecast.dev](mailto:mentoría@pullrecast.dev)