

Agentes de IA e o Futuro da tecnologia

Alan Braz

Lead Research Developer - IBM

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



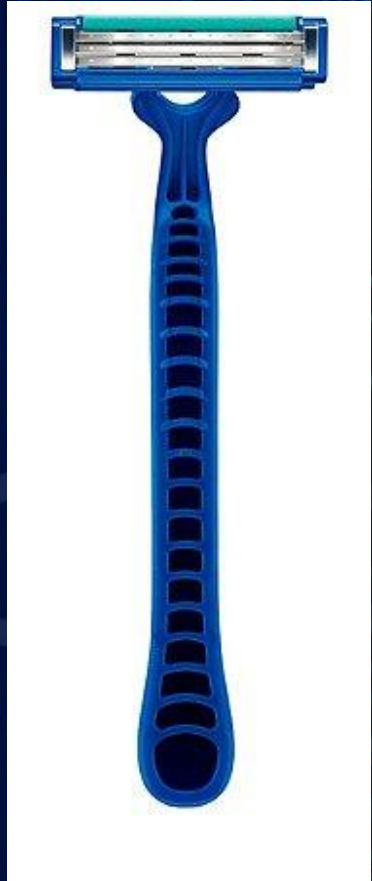
The Google logo is centered in the upper half of a white rounded rectangle. It consists of the word "Google" in its signature multi-colored font: blue 'G', red 'o', yellow 'o', blue 'g', green 'l', and red 'e'.

Google

Pesquise ou digite um URL



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: it won't answer questions about events after 2021

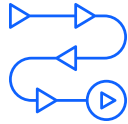
O melhor da IA

Quando usar

Recursos de IA Tradicional



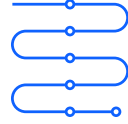
Preditivo/Prescritivo



IA conversacional dirigida



IA de Visão Dirigida



Automação de Processos



Quando usar

Recursos de IA Generativa



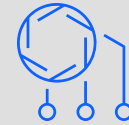
Sumarização



Pesquisa conversacional



Criação de conteúdo

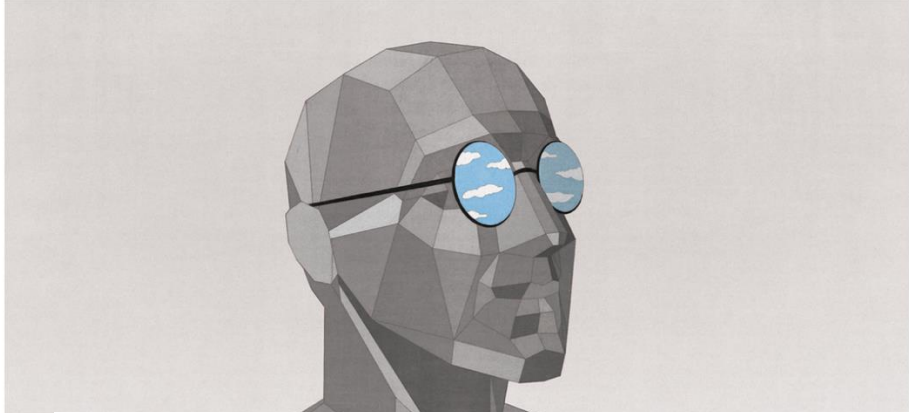


Criação de código

How People Are Really Using Gen AI in 2025

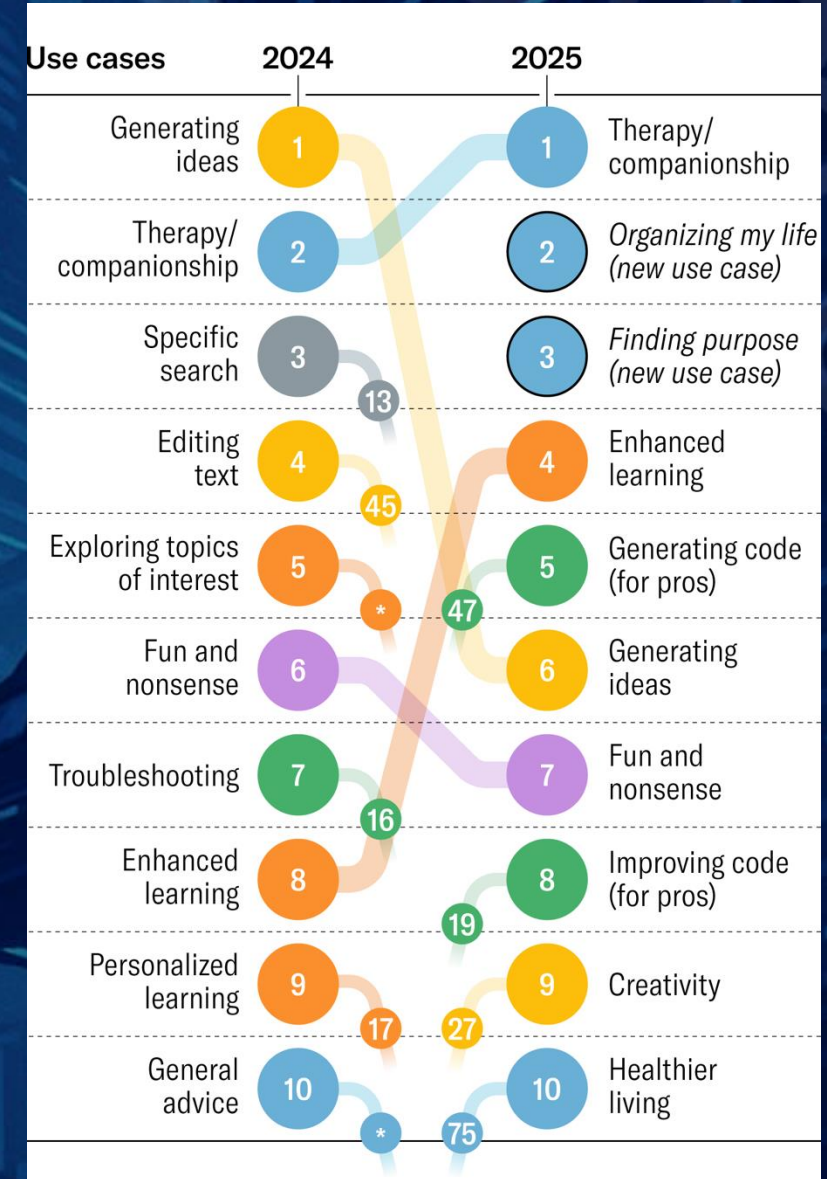
by Marc Zao-Sanders

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/>

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



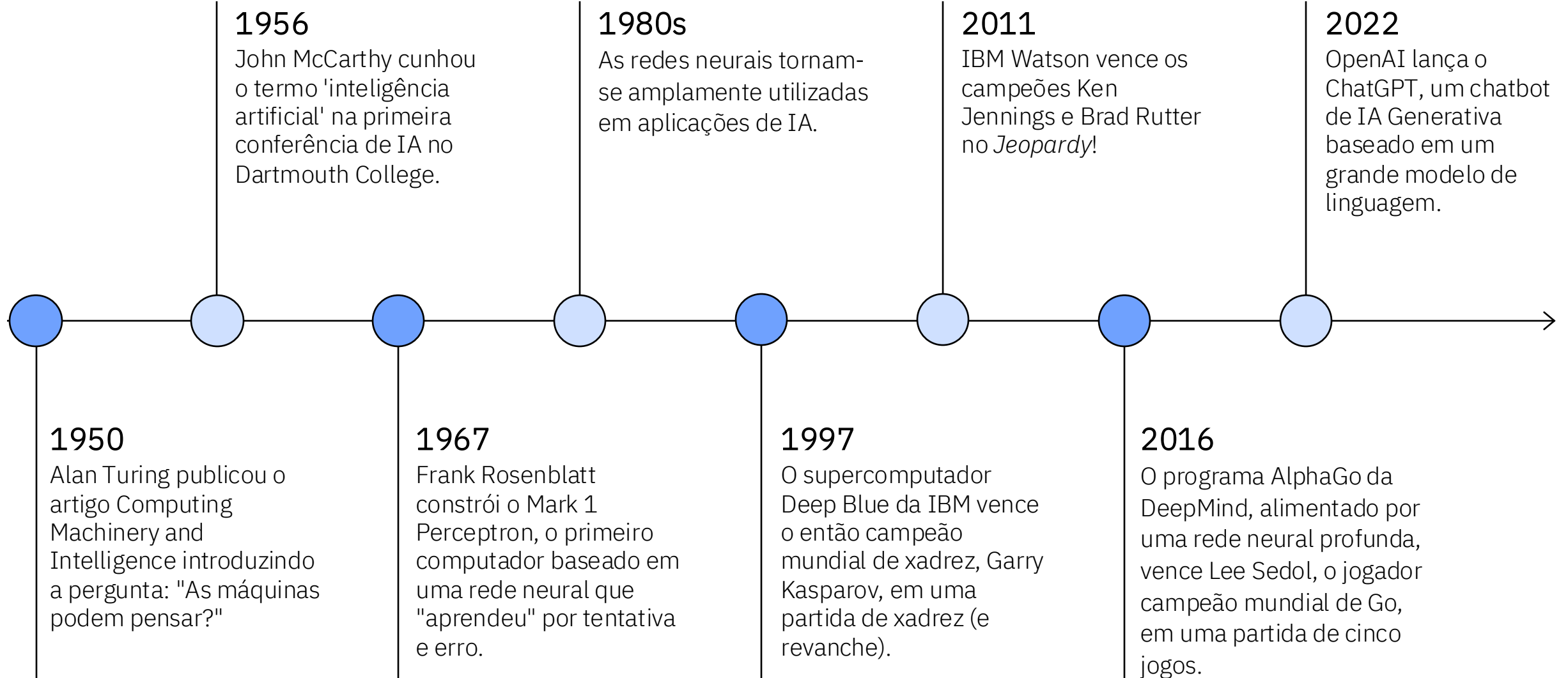
POR VINNY MATHIAS

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>



Marcos da IA



Deep Blue 1997



Ei! 2013/2014



2019

O Boticário lança 1's 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)

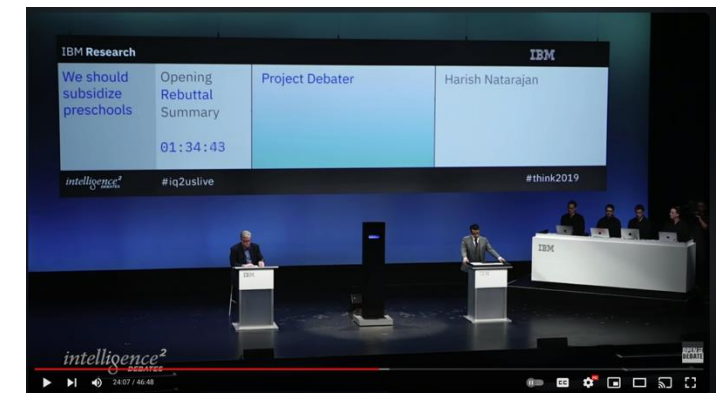
Watson 2011



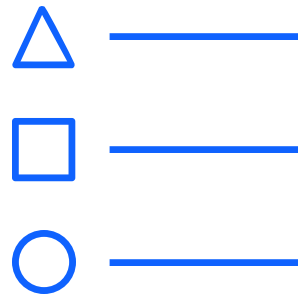
Chef Watson 2017



Project Debater



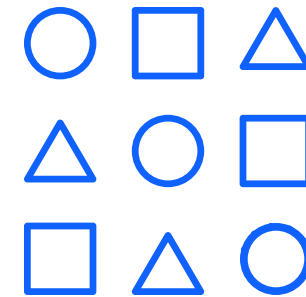
Sistemas baseados em regras



*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



Os sistemas baseados em ML são treinados em dados históricos para descobrir padrões. Os usuários fornecem entradas para o sistema de ML, que aplica essas entradas aos padrões descobertos e gera as saídas correspondentes.

Deep Learning (DL)

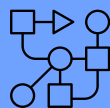
Técnica de ML que imita a função cerebral humana



DL é um subconjunto de ML, usando várias camadas de redes neurais, que são nós interconectados, que trabalham juntos para processar informações. O DL é adequado para aplicações complexas, como reconhecimento de imagem e fala.

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



82%

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

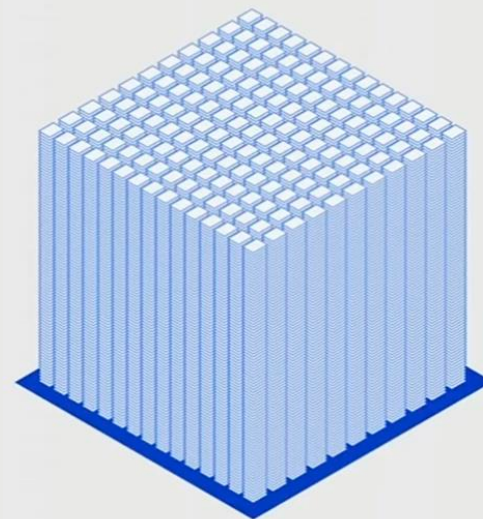
55%

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.



Transparente

Aberta à inspeção e avaliação.



Explicável

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



Justa

Treinada em dados imparciais.



Robusta

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



Privada

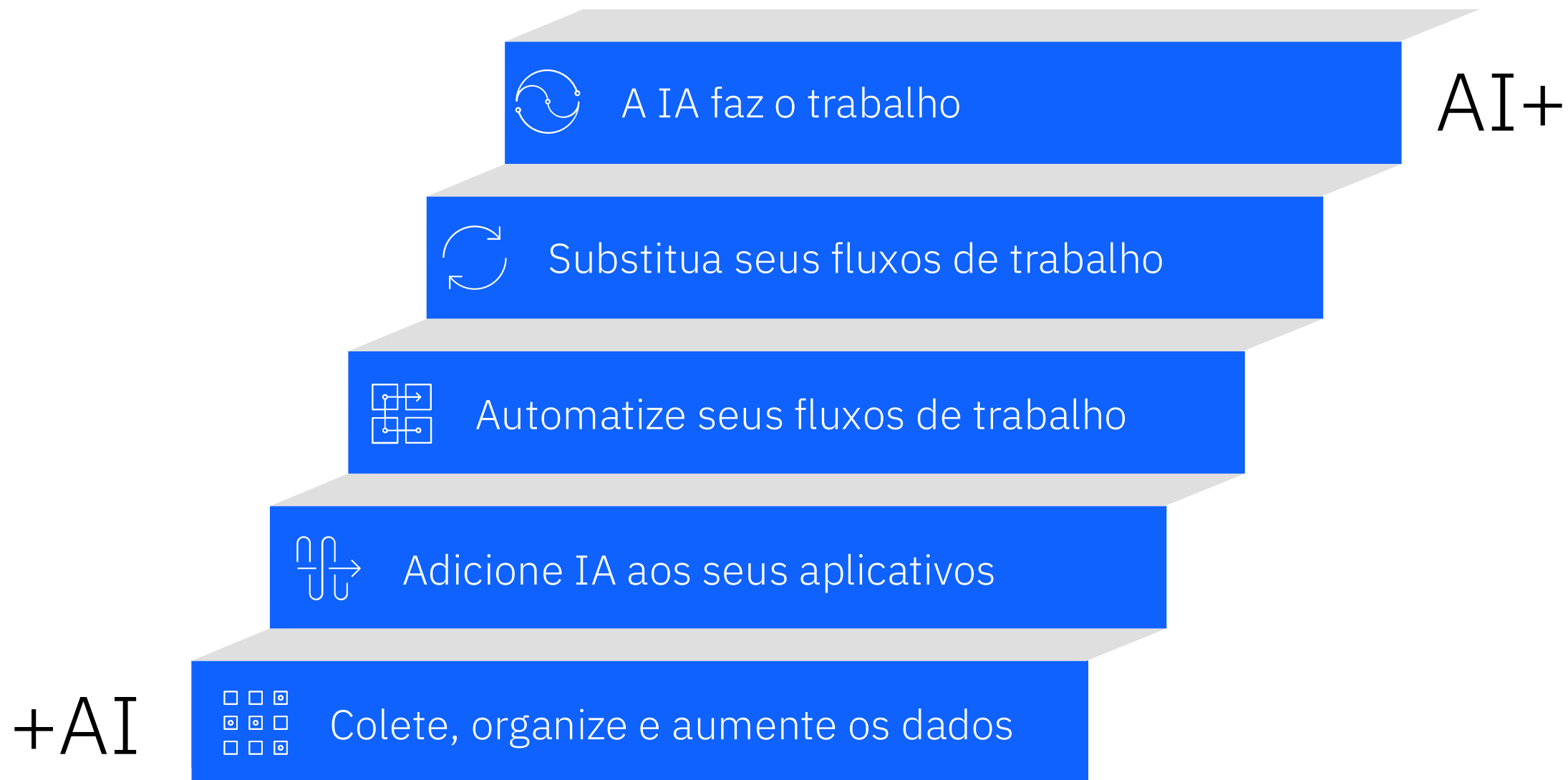
Alimentada por dados de alta integridade compatíveis com os negócios e defende a privacidade e necessidades de segurança.

EFICIENTE

+AI → AI+

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

A escada da IA moderna



Employee Experience ↻



Um único canal:
experiência
envolvente

2 camadas: Digital &
Humano



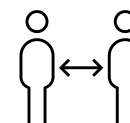
4,7k páginas
de conteúdo de RH



50+
Idiomas suportados



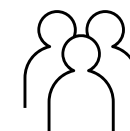
243k
Usuários
únicos



1.9m
Interações



97%
Adoção de gerentes



94%
Adoção de
executivos

Profissionais de RH atuando em atividades estratégicas

75%

Melhora em
produtividade

61%

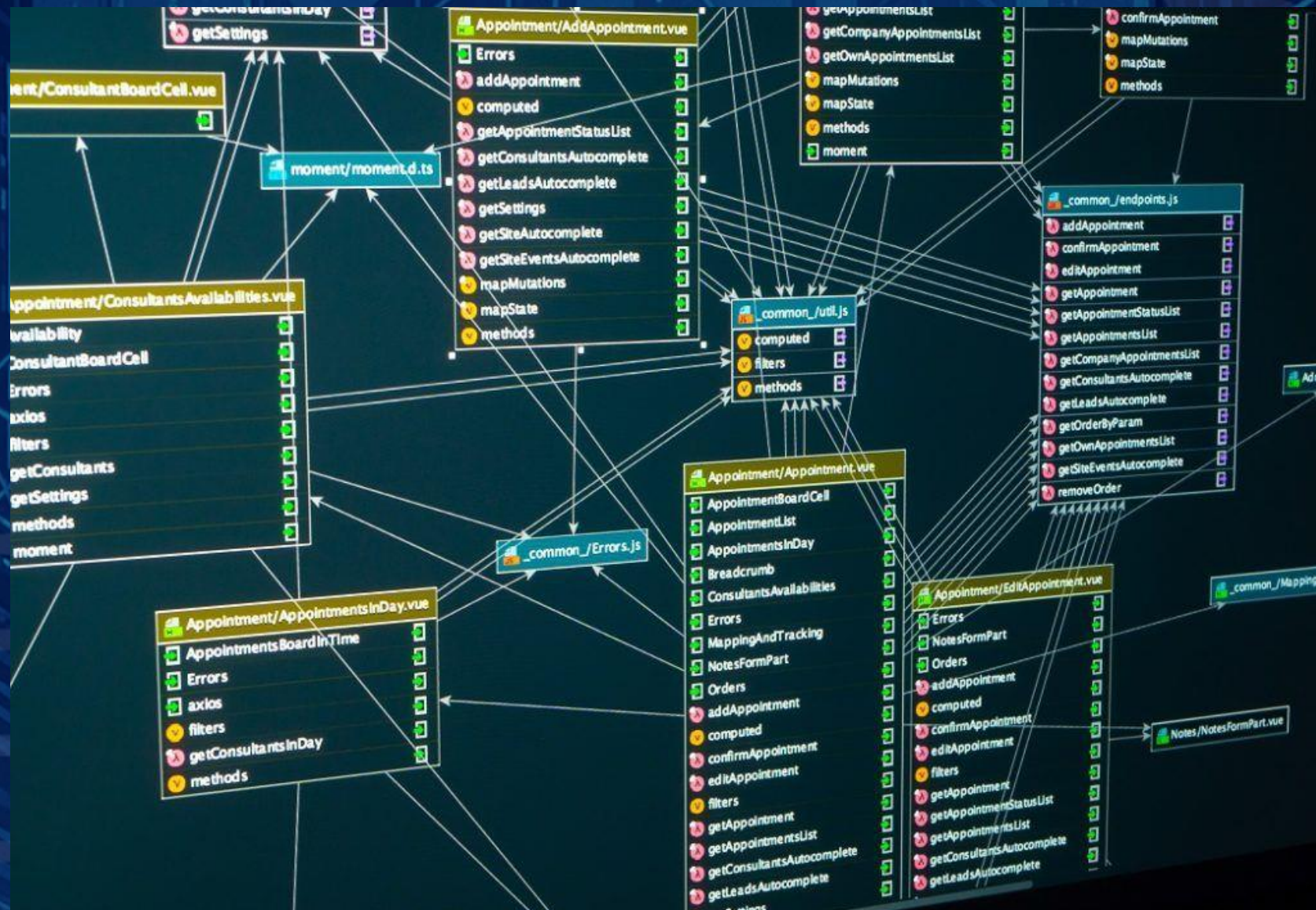
Redução de
abertura de
chamados

75%

Feedbacks
positivos
(CSAT)

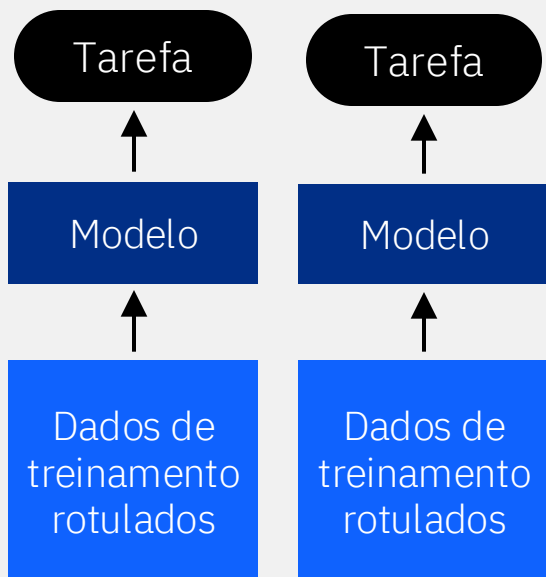
89%

de interações
resolvidas
diretamente via
AskHR



IA convencional

1000s – 100000s Pontos de dados rotulados / Tarefa



O que vem a seguir com os modelos de fundação:

Zero-shot prompting

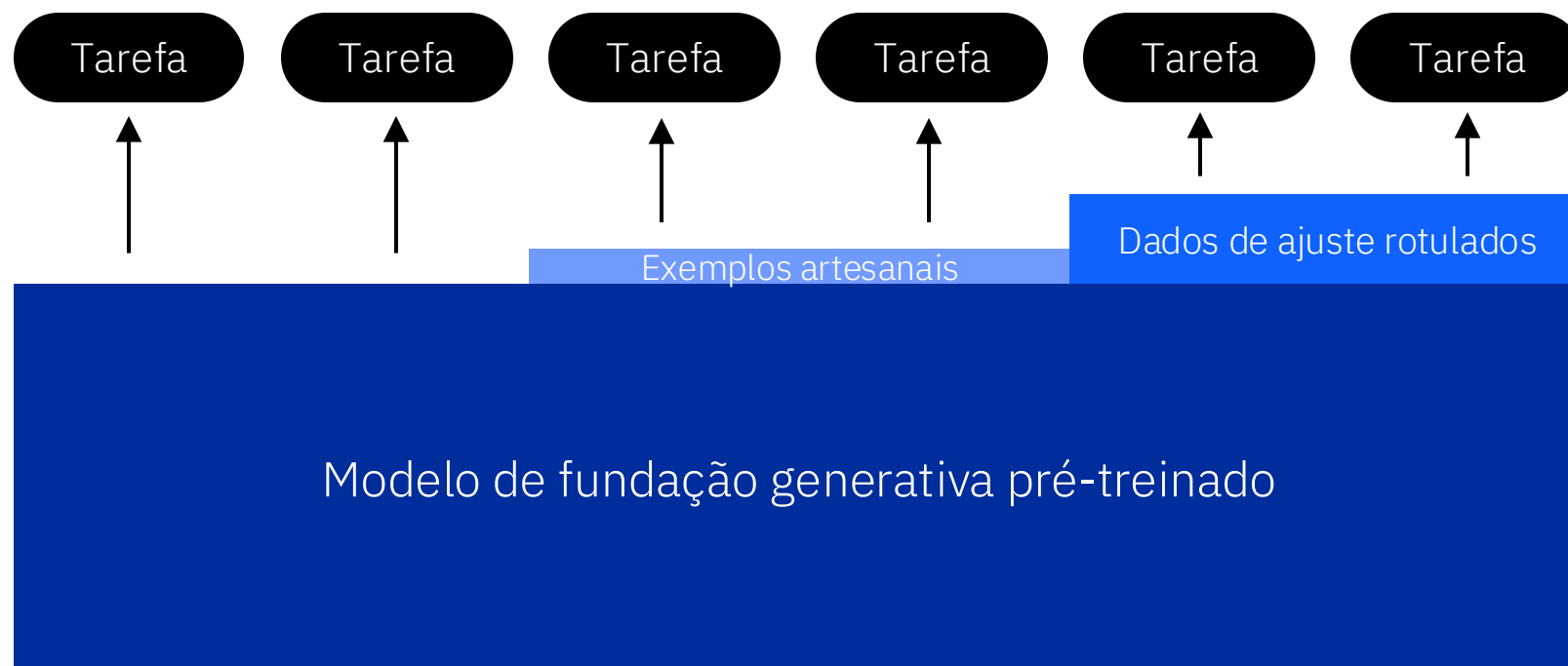
No labeled data needed

Few-shot prompting

1-10 labeled data points / task

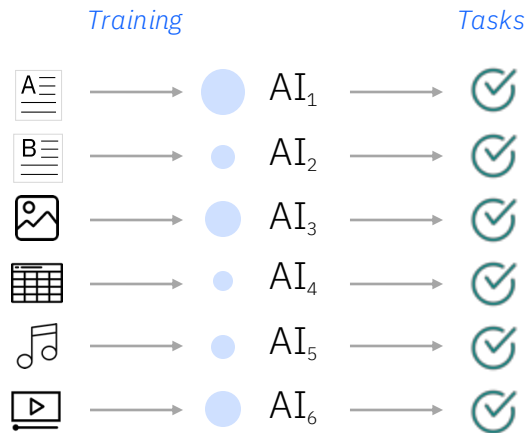
Prompt-tuning

100s – 1000s labeled data points / task



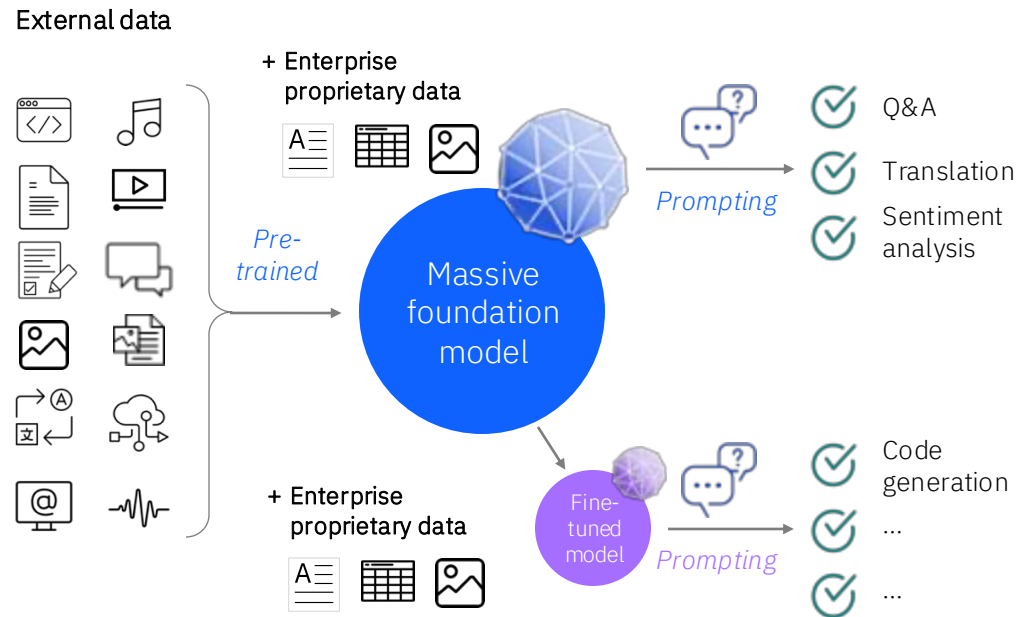
Foundation models establish a new paradigm for AI capabilities

Traditional AI models



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

Foundation models



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

Enhanced capabilities

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

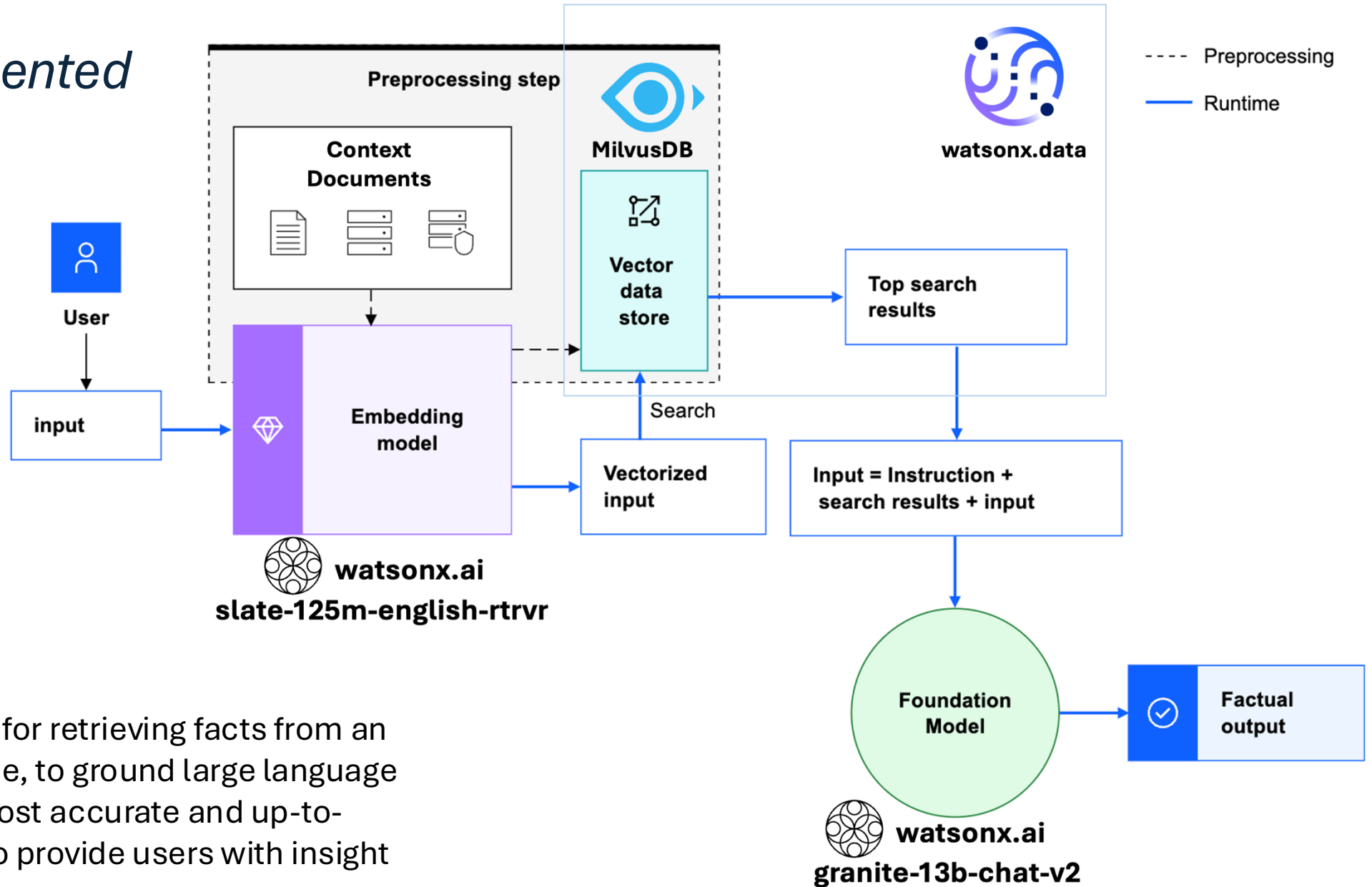
Key advantages

- 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

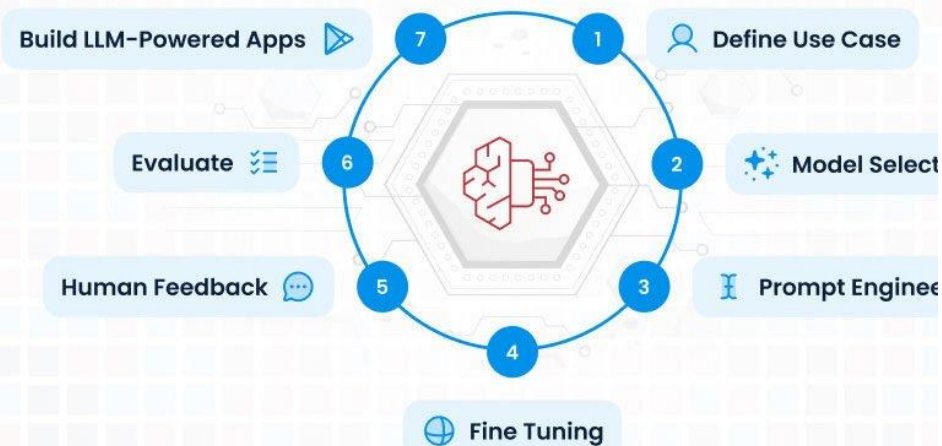
RAG

Retrieval Augmented Generation

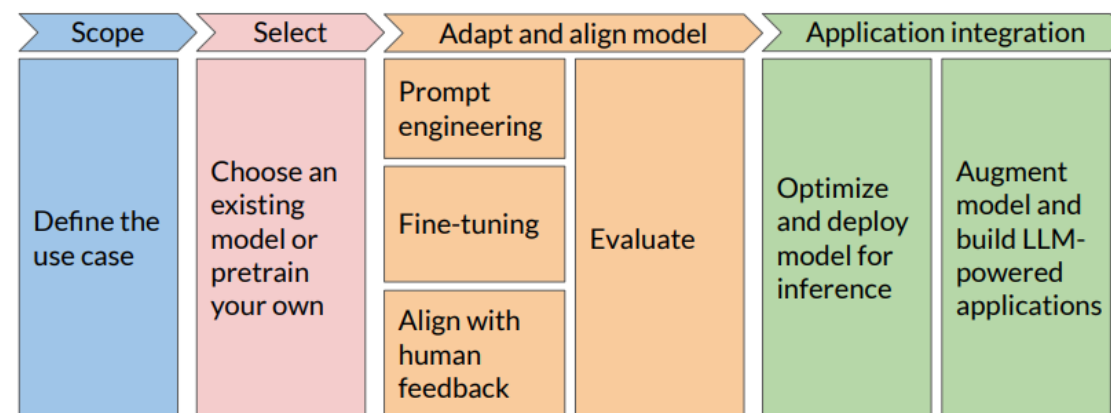


RAG is an AI framework for retrieving facts from an external knowledge base, to ground large language models (LLMs) in the most accurate and up-to-date information, and to provide users with insight into the LLM generating process.

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



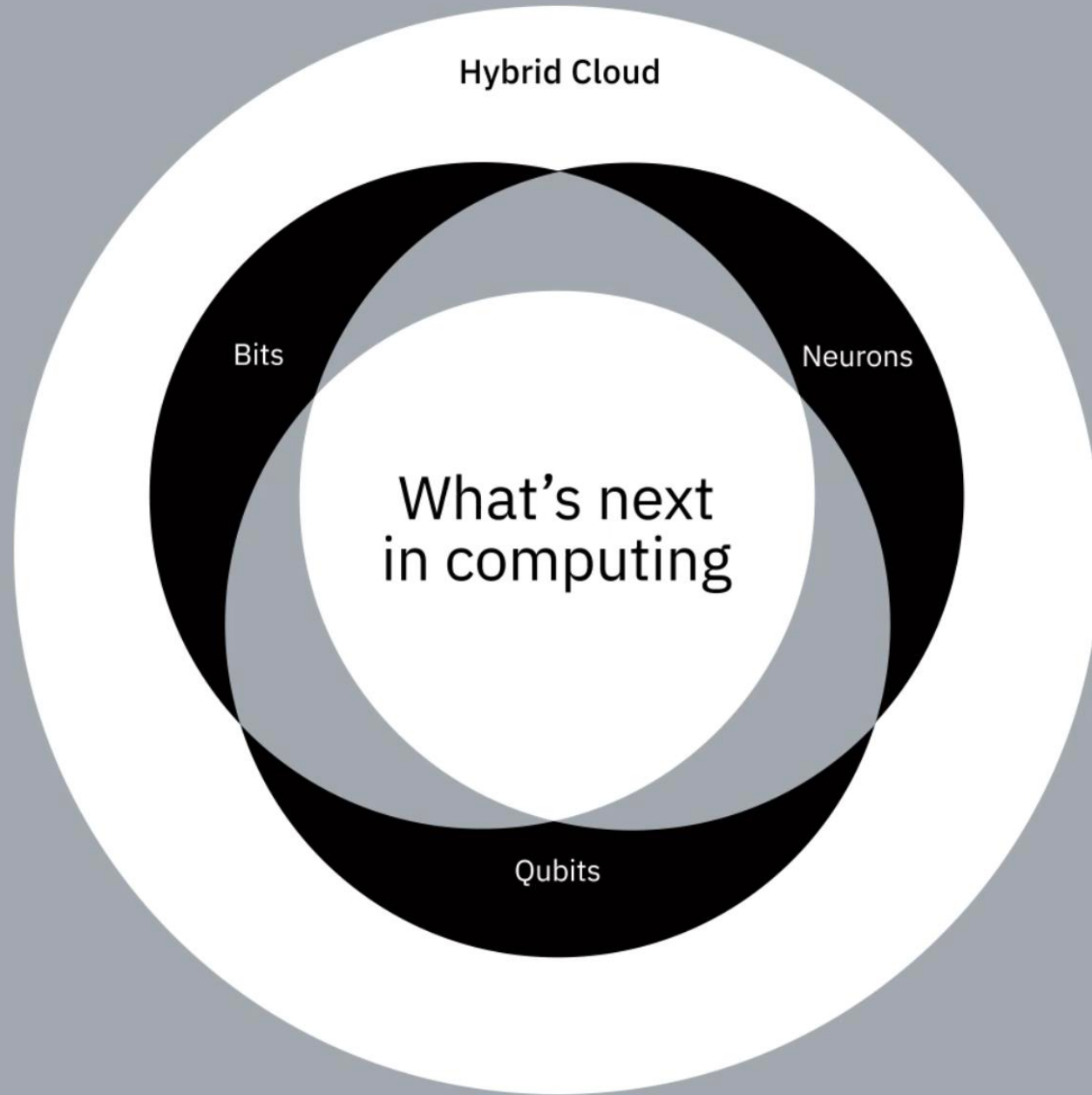
10 Medals of Technology



6 Turing Awards







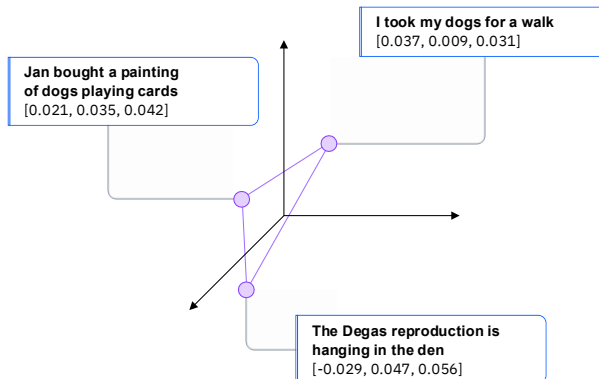
~~watsonx~~ AI Platform Engineering

- 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](#)

129 public models found.

Filters:

Lifecycle Stages

Readiness

Domain

Frameworks

☐ Computer Vision

☐ Graphs

☒ Medical (129)

☐ Miscellaneous

☐ Natural Language Processing

☐ Playing Games

☐ Speech

☐ Unknown

☐ 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%) Training (30%)	Trainability (50%)	16 Aug 2019
HoVer-Net: Simultaneous Segmentation and Classification of Nuclei in Multi-Tissue Histology Images ★ 17	Medical	TensorFlow	Definition (70%) Training (30%)	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 (https://arxiv.org/abs/1605.05273) ★ 19	Medical	TensorFlow	Definition (70%) Training (30%)	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%) Training (30%)	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%) Training (30%)	Trainability (50%)	16 Aug 2019
Complementary_Segmentation_Network (Outperforms u-nets everytime :) for binary segmentation) ★ 15	Medical	TensorFlow, Keras	Definition (70%) Training (30%)	Trainability (50%)	16 Aug 2019
Complementary_Segmentation_Network (Outperforms u-nets everytime :) for binary segmentation) ★ 15	Medical	TensorFlow, Keras	Definition (70%) Training (30%)	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**

Search models, datasets, users...

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_solubi...

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_bi...

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

7.09k

LTX Video Fast

ultra-fast video model, LTX 0.9.7 13B distilled

461

google/medgemma-4b-it

Image-Text-to-Text • Updated 3 d... • 8.49k • 172

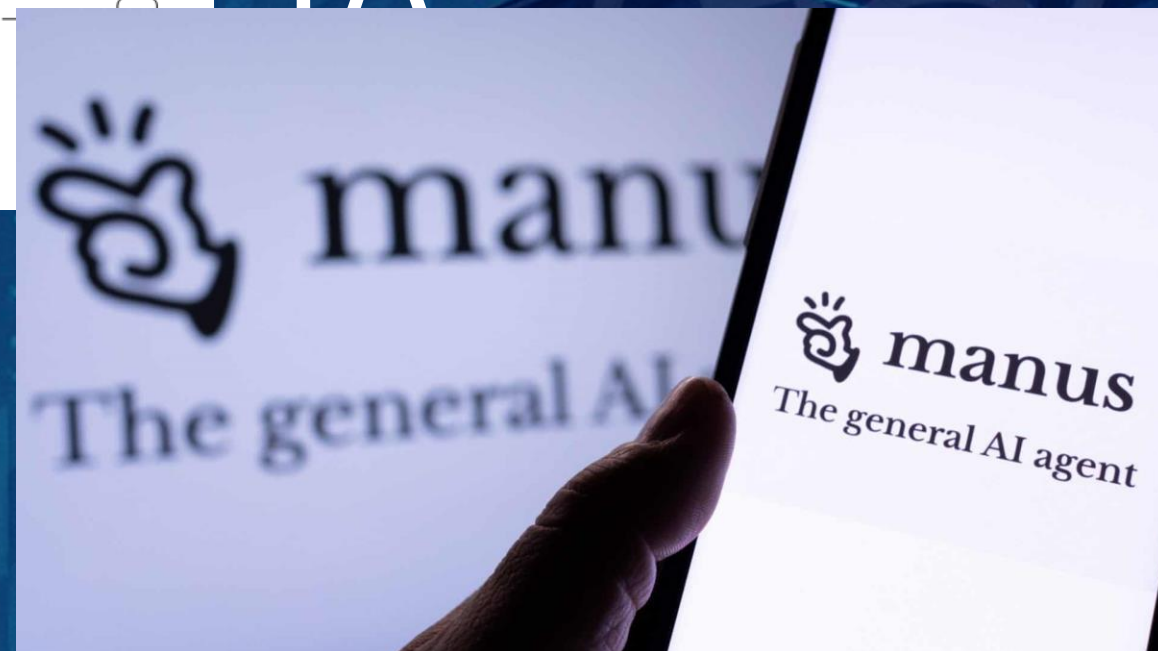
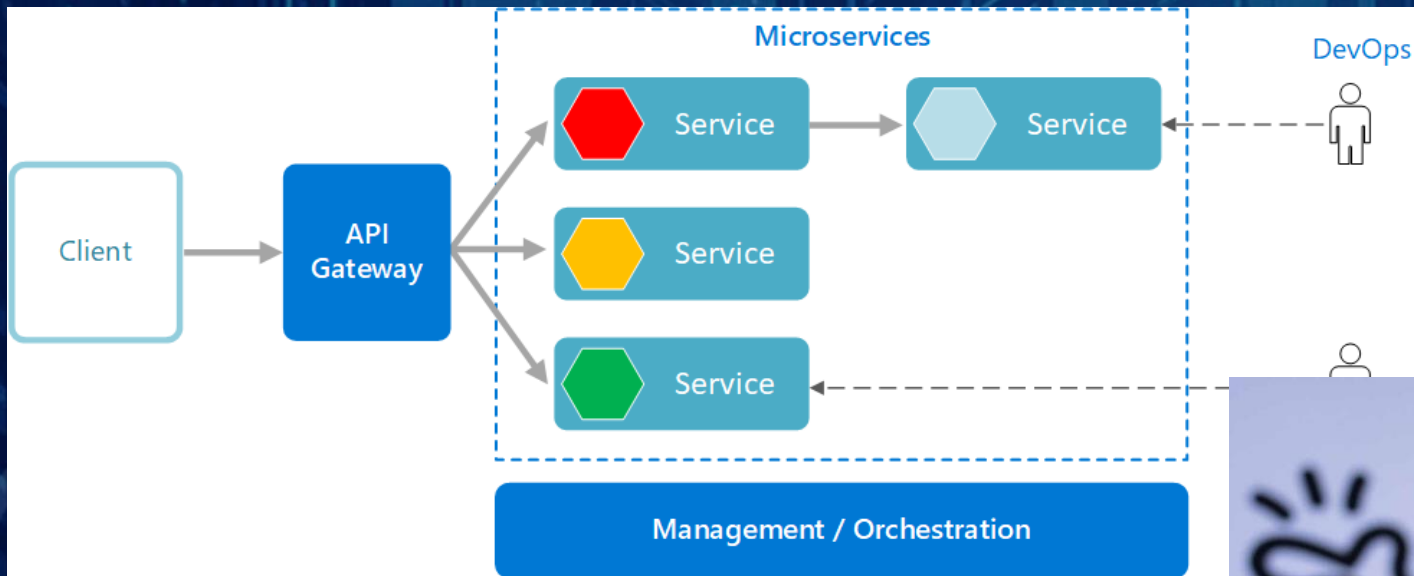
FLUX Pro Unlimited

Use the FLUX-Pro model as much as you want.

643

disco-eth/EuroSpeech

Viewer • Updated 6 day... • 8.42M • 27.4k • 52



As 10 principais tendências tecnológicas estratégicas para 2025



Imperativos e riscos da IA

- IA agêntica
- Plataformas de governança de IA
- Segurança contra desinformação



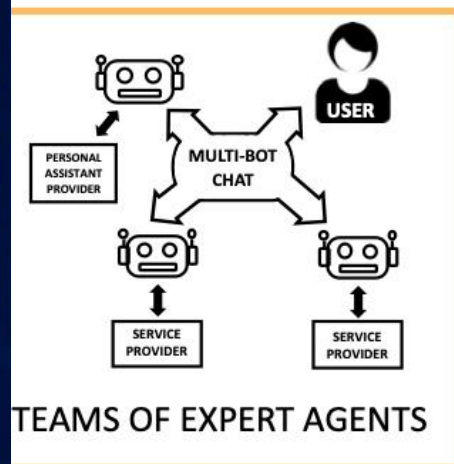
Novas fronteiras da computação

- Criptografia pós-quântica
- Inteligência ambiental invisível
- Computação verde
- Computação híbrida



Sinergia entre humanos e máquinas

- Computação espacial
- Robôs polifuncionais
- Aprimoramento neurológico



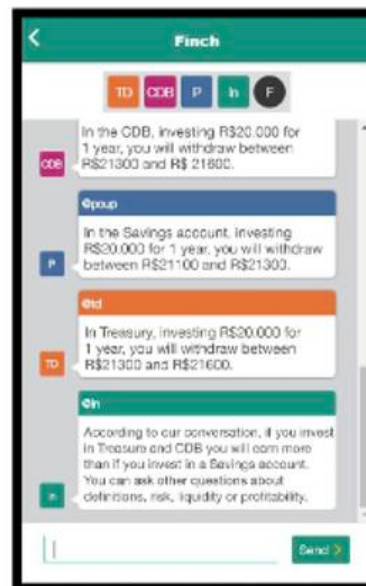
cdbGuru
reactive bot



poupancaGuru
reactive bot



tesouroGuru
reactive bot



from Q&A dialog to multiparty chat



User



investmentGuru
proactive bot

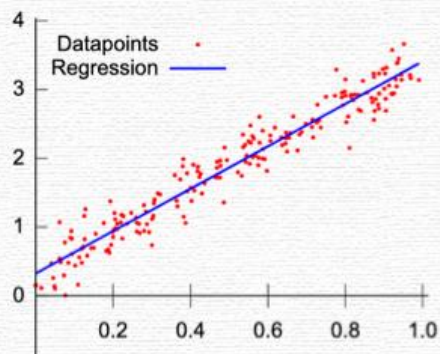
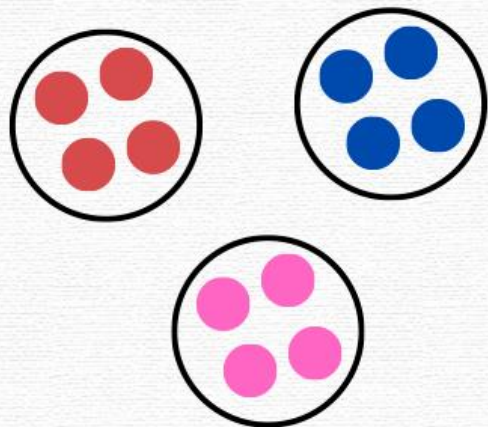
15

Self-service finance advising
using multiple chatbots

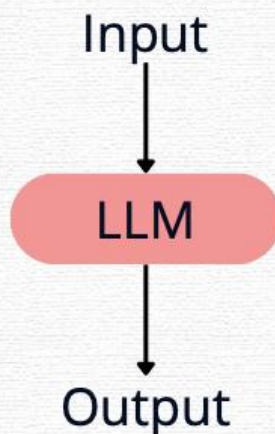
Should we design multi or single bots?
[DeepDial'18]

Evolução da Inteligência Artificial

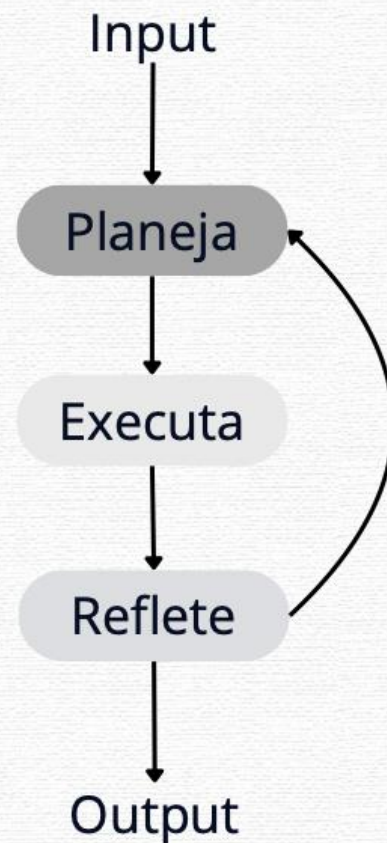
➤ Classificação/Predição



➤ Assistants



➤ AI Agent

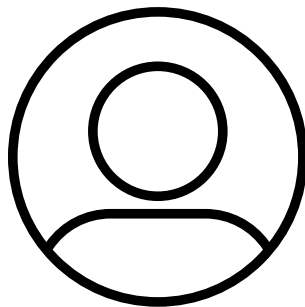


Introdução

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

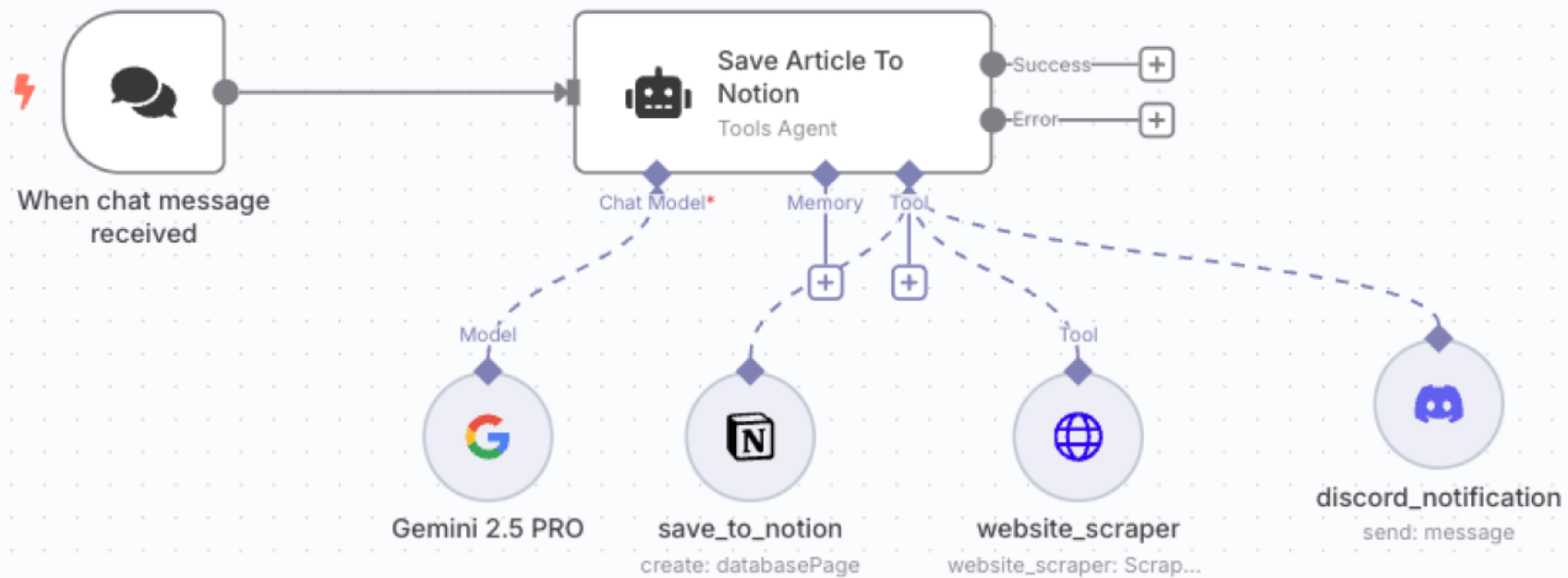
Agent: (tool_caption) Getting current weather for Las Vegas

```
> 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, "generation_time_ms": 0.08392333984375, "utc_offset_seconds": 0, "timezone": "UTC", "timezone_abbrev": "UTC", "units": "seconds", "temperature_2m": "°C", "rain": "mm", "apparent_temperature": "°C"}, "current": {"time": "2024-09-13T11:15", "interval": 900, "temperature_2m": 22.3, "rain": 0, "apparent_temperature": 22.3, "relative_humidity_2m": "%", "apparent_temperature": "°C"}, "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, 22.3, 21.8, 21.3, 20.8, 20.3, 19.8, 19.3, 18.8, 18.3, 17.8, 17.3, 16.8, 16.3, 15.8, 15.3, 14.8, 14.3, 13.8, 13.3, 12.8, 12.3, 11.8, 11.3, 10.8, 10.3, 9.8, 9.3, 8.8, 8.3, 7.8, 7.3, 6.8, 6.3, 5.8, 5.3, 4.8, 4.3, 3.8, 3.3, 2.8, 2.3, 1.8, 1.3, 0.8, 0.3, -0.2, -0.7, -1.2, -1.7, -2.2, -2.7, -3.2, -3.7, -4.2, -4.7, -5.2, -5.7, -6.2, -6.7, -7.2, -7.7, -8.2, -8.7, -9.2, -9.7, -10.2, -10.7, -11.2, -11.7, -12.2, -12.7, -13.2, -13.7, -14.2, -14.7, -15.2, -15.7, -16.2, -16.7, -17.2, -17.7, -18.2, -18.7, -19.2, -19.7, -20.2, -20.7, -21.2, -21.7, -22.2, -22.7, -23.2, -23.7, -24.2, -24.7, -25.2, -25.7, -26.2, -26.7, -27.2, -27.7, -28.2, -28.7, -29.2, -29.7, -30.2, -30.7, -31.2, -31.7, -32.2, -32.7, -33.2, -33.7, -34.2, -34.7, -35.2, -35.7, -36.2, -36.7, -37.2, -37.7, -38.2, -38.7, -39.2, -39.7, -40.2, -40.7, -41.2, -41.7, -42.2, -42.7, -43.2, -43.7, -44.2, -44.7, -45.2, -45.7, -46.2, -46.7, -47.2, -47.7, -48.2, -48.7, -49.2, -49.7, -50.2, -50.7, -51.2, -51.7, -52.2, -52.7, -53.2, -53.7, -54.2, -54.7, -55.2, -55.7, -56.2, -56.7, -57.2, -57.7, -58.2, -58.7, -59.2, -59.7, -60.2, -60.7, -61.2, -61.7, -62.2, -62.7, -63.2, -63.7, -64.2, -64.7, -65.2, -65.7, -66.2, -66.7, -67.2, -67.7, -68.2, -68.7, -69.2, -69.7, -70.2, -70.7, -71.2, -71.7, -72.2, -72.7, -73.2, -73.7, -74.2, -74.7, -75.2, -75.7, -76.2, -76.7, -77.2, -77.7, -78.2, -78.7, -79.2, -79.7, -80.2, -80.7, -81.2, -81.7, -82.2, -82.7, -83.2, -83.7, -84.2, -84.7, -85.2, -85.7, -86.2, -86.7, -87.2, -87.7, -88.2, -88.7, -89.2, -89.7, -90.2, -90.7, -91.2, -91.7, -92.2, -92.7, -93.2, -93.7, -94.2, -94.7, -95.2, -95.7, -96.2, -96.7, -97.2, -97.7, -98.2, -98.7, -99.2, -99.7, -100.2, -100.7, -101.2, -101.7, -102.2, -102.7, -103.2, -103.7, -104.2, -104.7, -105.2, -105.7, -106.2, -106.7, -107.2, -107.7, -108.2, -108.7, -109.2, -109.7, -110.2, -110.7, -111.2, -111.7, -112.2, -112.7, -113.2, -113.7, -114.2, -114.7, -115.2, -115.7, -116.2, -116.7, -117.2, -117.7, -118.2, -118.7, -119.2, -119.7, -120.2, -120.7, -121.2, -121.7, -122.2, -122.7, -123.2, -123.7, -124.2, -124.7, -125.2, -125.7, -126.2, -126.7, -127.2, -127.7, -128.2, -128.7, -129.2, -129.7, -130.2, -130.7, -131.2, -131.7, -132.2, -132.7, -133.2, -133.7, -134.2, -134.7, -135.2, -135.7, -136.2, -136.7, -137.2, -137.7, -138.2, -138.7, -139.2, -139.7, -140.2, -140.7, -141.2, -141.7, -142.2, -142.7, -143.2, -143.7, -144.2, -144.7, -145.2, -145.7, -146.2, -146.7, -147.2, -147.7, -148.2, -148.7, -149.2, -149.7, -150.2, -150.7, -151.2, -151.7, -152.2, -152.7, -153.2, -153.7, -154.2, -154.7, -155.2, -155.7, -156.2, -156.7, -157.2, -157.7, -158.2, -158.7, -159.2, -159.7, -160.2, -160.7, -161.2, -161.7, -162.2, -162.7, -163.2, -163.7, -164.2, -164.7, -165.2, -165.7, -166.2, -166.7, -167.2, -167.7, -168.2, -168.7, -169.2, -169.7, -170.2, -170.7, -171.2, -171.7, -172.2, -172.7, -173.2, -173.7, -174.2, -174.7, -175.2, -175.7, -176.2, -176.7, -177.2, -177.7, -178.2, -178.7, -179.2, -179.7, -180.2, -180.7, -181.2, -181.7, -182.2, -182.7, -183.2, -183.7, -184.2, -184.7, -185.2, -185.7, -186.2, -186.7, -187.2, -187.7, -188.2, -188.7, -189.2, -189.7, -190.2, -190.7, -191.2, -191.7, -192.2, -192.7, -193.2, -193.7, -194.2, -194.7, -195.2, -195.7, -196.2, -196.7, -197.2, -197.7, -198.2, -198.7, -199.2, -199.7, -200.2, -200.7, -201.2, -201.7, -202.2, -202.7, -203.2, -203.7, -204.2, -204.7, -205.2, -205.7, -206.2, -206.7, -207.2, -207.7, -208.2, -208.7, -209.2, -209.7, -210.2, -210.7, -211.2, -211.7, -212.2, -212.7, -213.2, -213.7, -214.2, -214.7, -215.2, -215.7, -216.2, -216.7, -217.2, -217.7, -218.2, -218.7, -219.2, -219.7, -220.2, -220.7, -221.2, -221.7, -222.2, -222.7, -223.2, -223.7, -224.2, -224.7, -225.2, -225.7, -226.2, -226.7, -227.2, -227.7, -228.2, -228.7, -229.2, -229.7, -230.2, -230.7, -231.2, -231.7, -232.2, -232.7, -233.2, -233.7, -234.2, -234.7, -
```

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

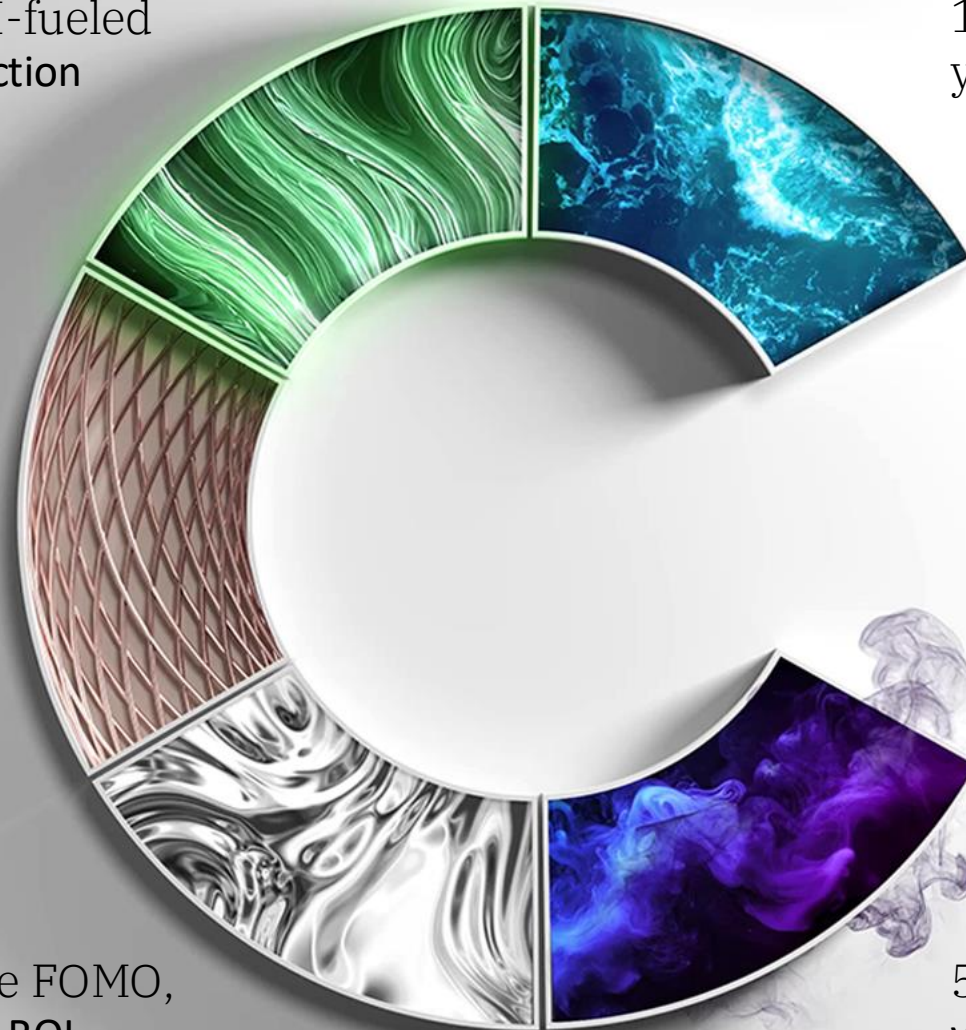
2. Embrace AI-fueled creative destruction

1. Make **courage** your core

3. Cultivate a vibrant data environment

4. Ignore FOMO, lean into ROI

5. Borrow the talent you can't buy



Download the full study
ibm.co/ceo-2025



“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



Amy Webb Launches 2025 Emerging Tech Trend Report | SXSW LIVE

96K views • Streamed 2 months ago



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

alanbraz.com.br / 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.





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-7330519792300322816-1yLW

alanbraz.com.br / pullrecast.dev



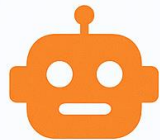
FUTURE TRENDS



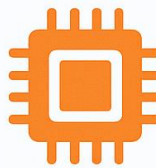
OpenSource
models



SLMs
Small Language



Agentic AI



Specific
hardware



AI Platforms



Quantum
computing



קוד פתוח
open source project



OpenSource Models



Chat with DeepSeek AI

 Meta
INTRODUCING
 LLaMA - 3



Gemma



Announcing

Phi-2

Built by
Microsoft Research

2.7 billion
parameters

Common sense,
language understanding,
and logical reasoning

aka.ms/Phi2

Alibaba.com



Qwen

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 Train Deploy Use this model

Edit model card

DeepSeek-R1



Downloads last month
735,368

Safetensors

Model size 685B params Tensor type BF16 · F8_E4M3 · F32

Chat template Files info

Inference Providers NEW

Hyperbolic +3

Text Generation Examples

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

Model size 685B params

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](#)[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 latest	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

deepseek

New chat

Today

H100 GPU Requirements for

7 Days

Explicação simples de inteligência

2025-04

Como remover linhas duplicadas

Alan Braz's Professional Profile S

Currículo de Pós-Graduação em C

2025-01

Meaning and Explanation of GSH

NBA Games in NY or Brooklyn Fel

Cardápio semanal saudável para f

Get App

My Profile

H100 GPU Requirements for DeepSeek R1

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

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

1. Memory Requirements (FP16/BF16)

- Parameters: 671 billion × 2 bytes = **1,342 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

4. Estimated GPU Count for 671B Model

Scenario	Minimum H100 GPUs	Recommended GPUs
Inference	17	20–24+
Training	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 VIDEO

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 WAZ

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

3x de R\$ 93.627,45 com 4.5% de desconto

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

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

[Ver mais](#)

Calcular o Frete

Digite seu CEP

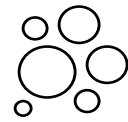
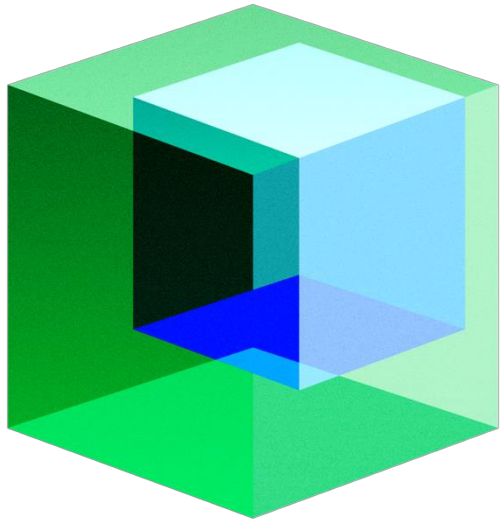
CALCULAR

[Não sei meu CEP](#)

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

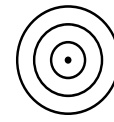
IBM Granite

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

huggingface.com/ibm-granite



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



Vision model

- Granite-Vision-3.2-2B



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 browser window with the URL `chat.alanbraz.com.br/c/d267d8d8-2979-4cf9-af99-e14a2ea14e9b`. The interface is dark-themed and includes a sidebar on the left with navigation options: New Chat, Notes, Workspace, Search, Chats, and a list of previous chats. The main chat area displays a conversation with the model `granite3-moe:1b`. The user's message is: "Como você explicaria inteligência artificial brevemente para minha avó de 89 anos?". The model's response is: "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 seniors e a gestão de processos. Em relação à sua idade, o melhor aconselhamento é fornecer informações detalhados 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." The interface also features a bottom bar with a "Send a Message" input, a "Web Search" button, a "Code Interpreter" button, and a microphone icon.

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

Alan Braz

granite3-moe:1b

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

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 seniors e a gestão de processos.

Em relação à sua idade, o melhor aconselhamento é fornecer informações detalhados 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.

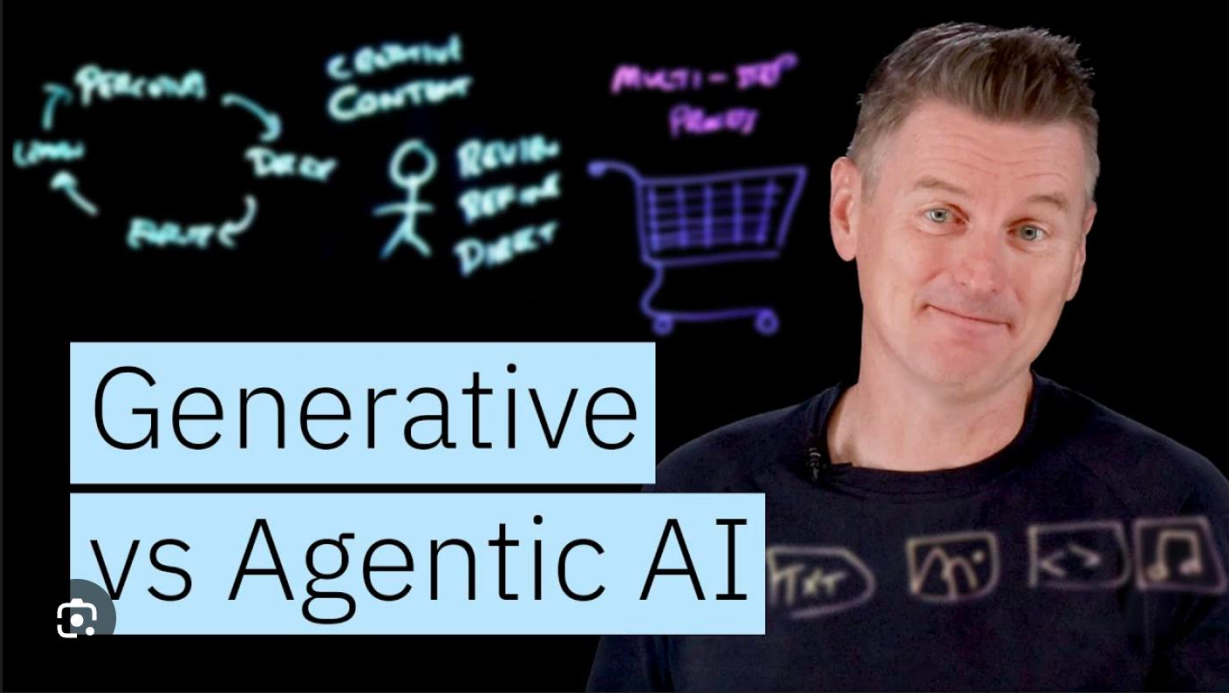
Send a Message

Web Search

Code Interpreter

YouTube
IBM Technology · 7:19

< > ⋮ ×



Generative vs Agentic AI

Generative vs Agentic AI: Shaping the Future of AI Collaboration

Assistir >

Data do upload: 21 de abr. de 2025 · 3,41 mil Curtidas

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



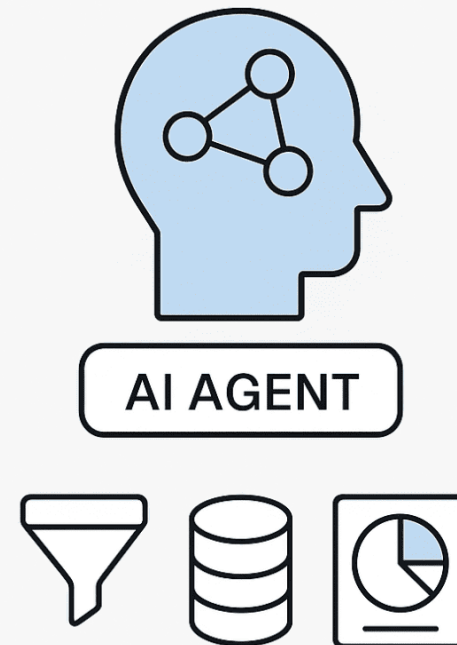
Agentic AI

Manage complex projects

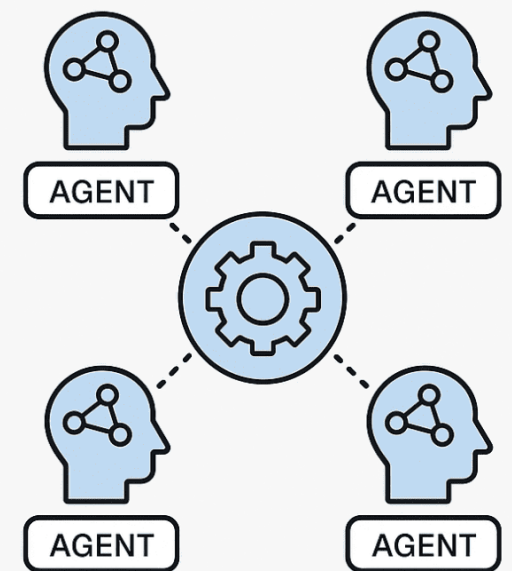
<https://www.mygreatlearning.com/blog/ai-agents-and-agentic-ai/>

<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

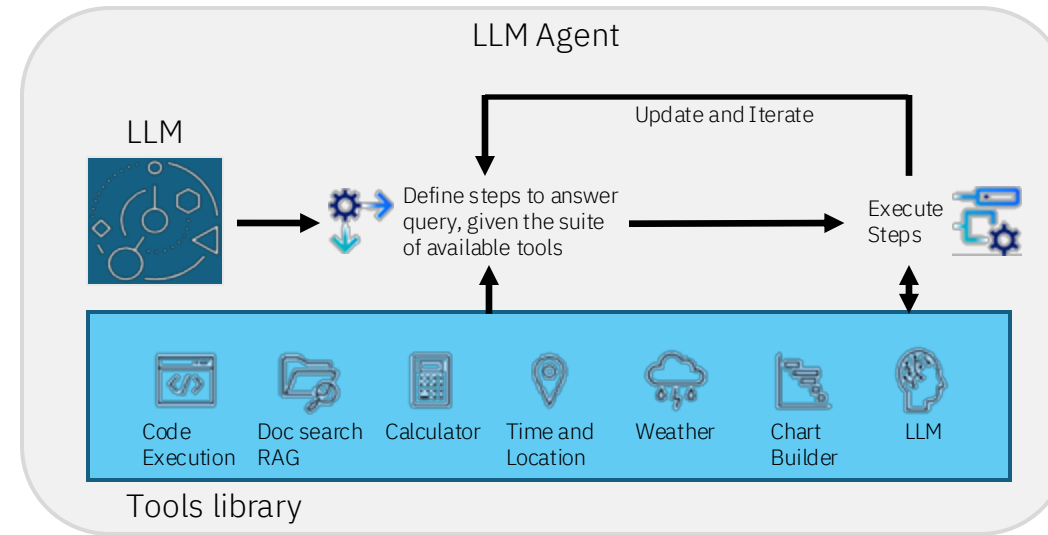


From Q&A to Agentic AI

Increasing use case complexities

Agentic AI

Incoming request



Multi-step processes
Autonomous action-taking
Self-correcting

Request outcome

RAG

Incoming query



Grounding information retrieval
Prescriptive tasks
Single-step processes

Output completion

LLM and Prompt Engineering

Incoming query

Zero/few-shot prompting



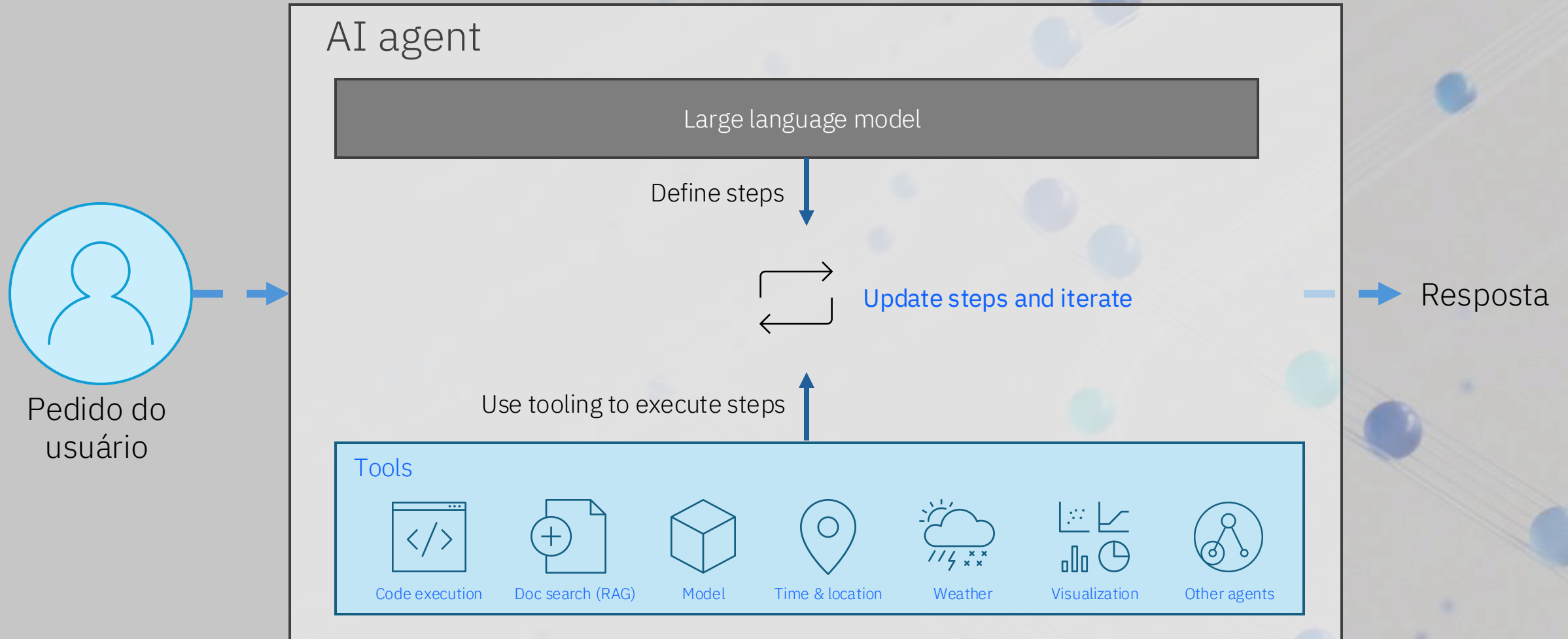
LLM

Problem-solving
Logical thinking
Pattern matching

Output completion

O que são AI Agents e como funcionam?

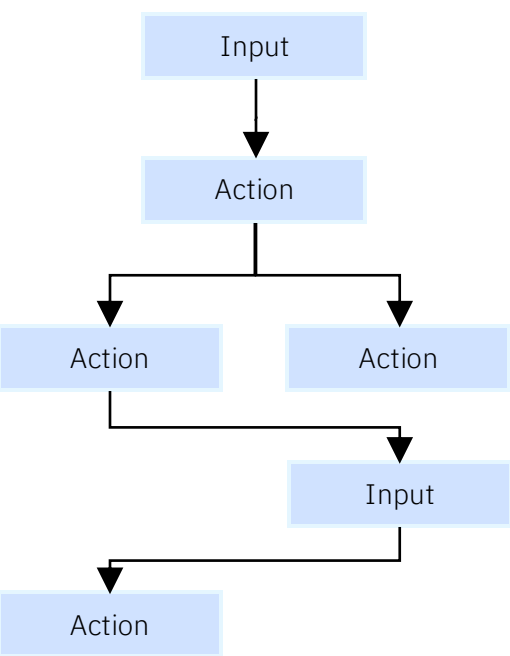
Um AI agent é um programa cuja lógica de execução é controlada por LLM



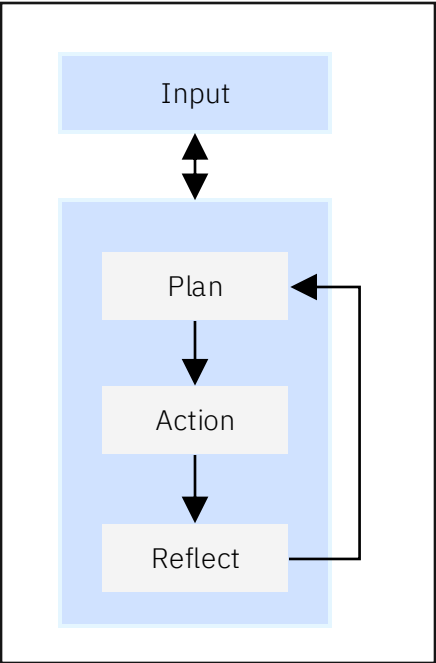
Do Assistente ao Multi-Agent

Assistentes tradicionais → Assistentes Single-agent → Assistentes Multi-agent

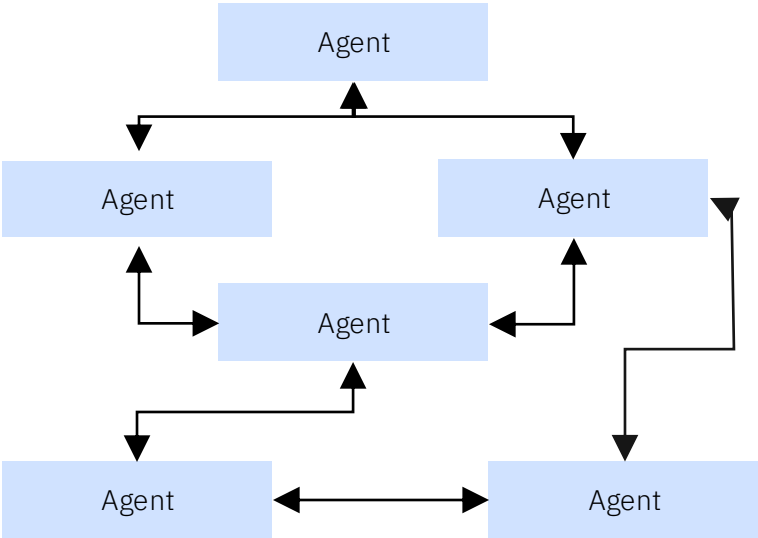
- Baseado em regras (se x, faça y)
- Caminhos de ação predefinidos



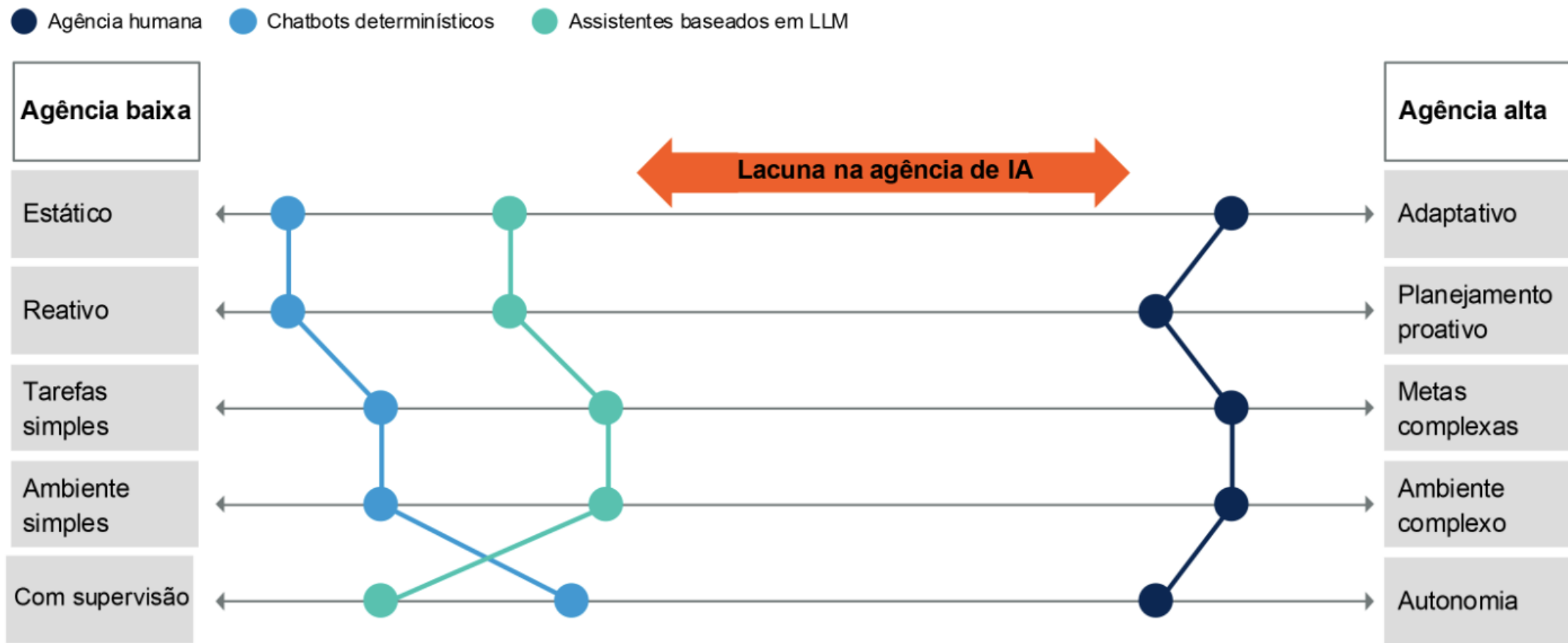
- Baseados em tarefas (ex.: reserva de voo)
- Restrições de desempenho conforme aumenta o escopo da tarefa
- Controle limitado



- 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

Gartner

Capacidades da IA estão crescendo rapidamente

IA que cria



Generative AI

IA que prediz



Machine learning

IA que conversa



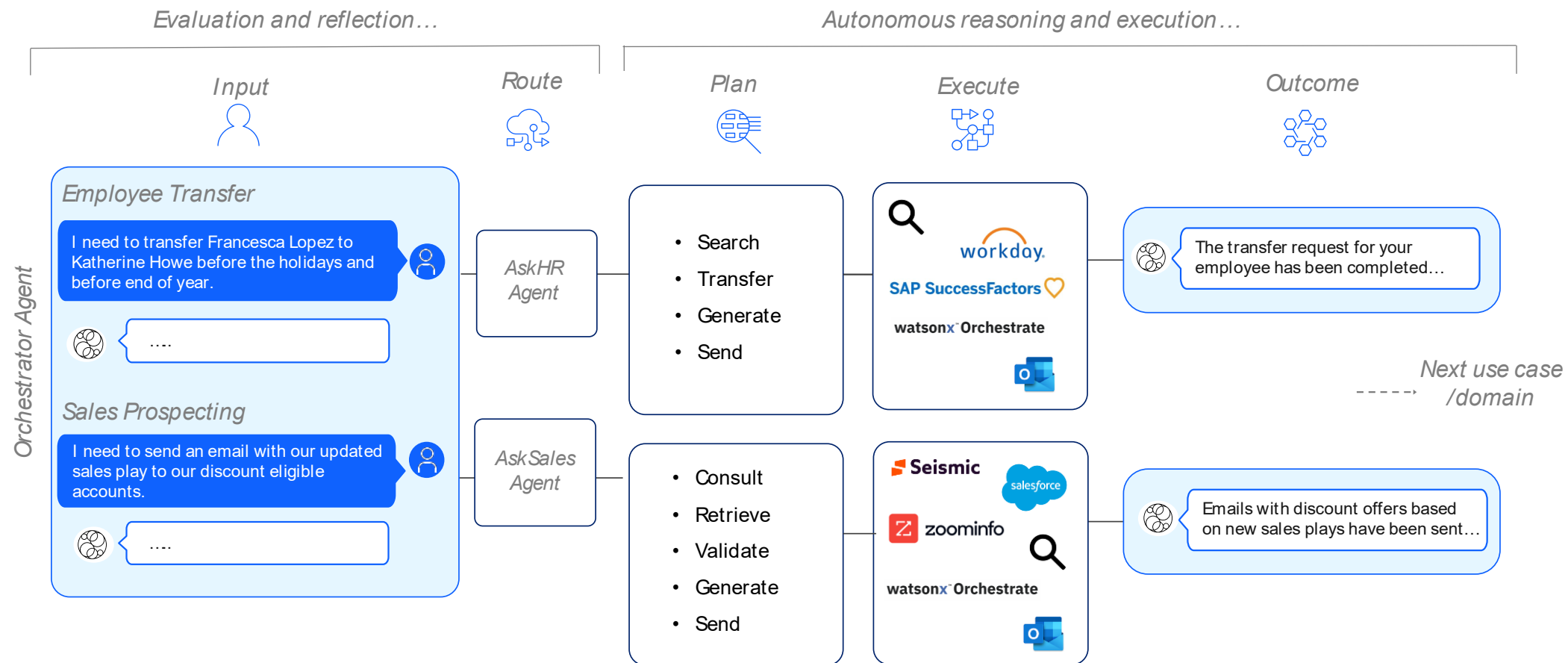
AI assistants

IA que trabalha

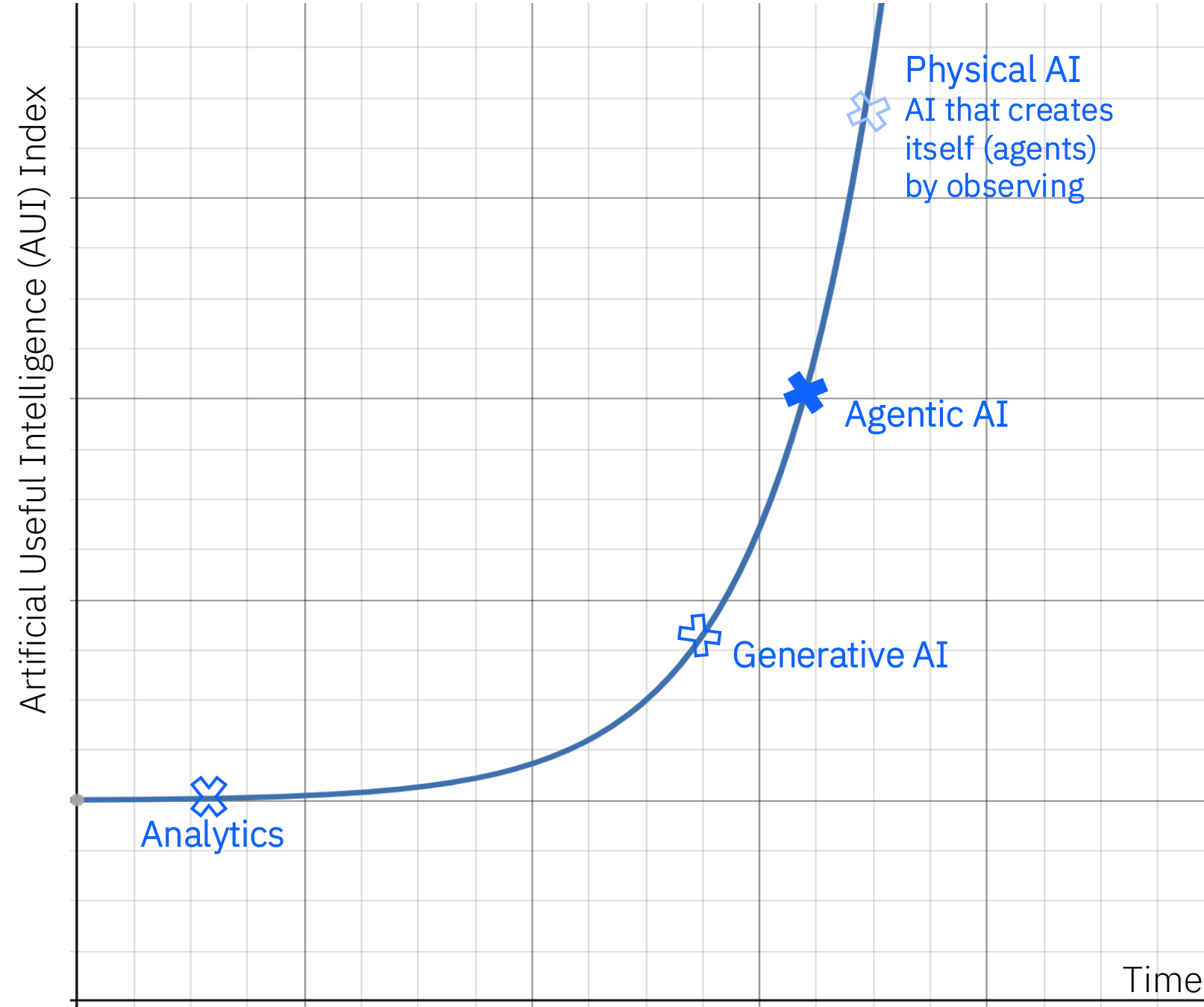


AI agents

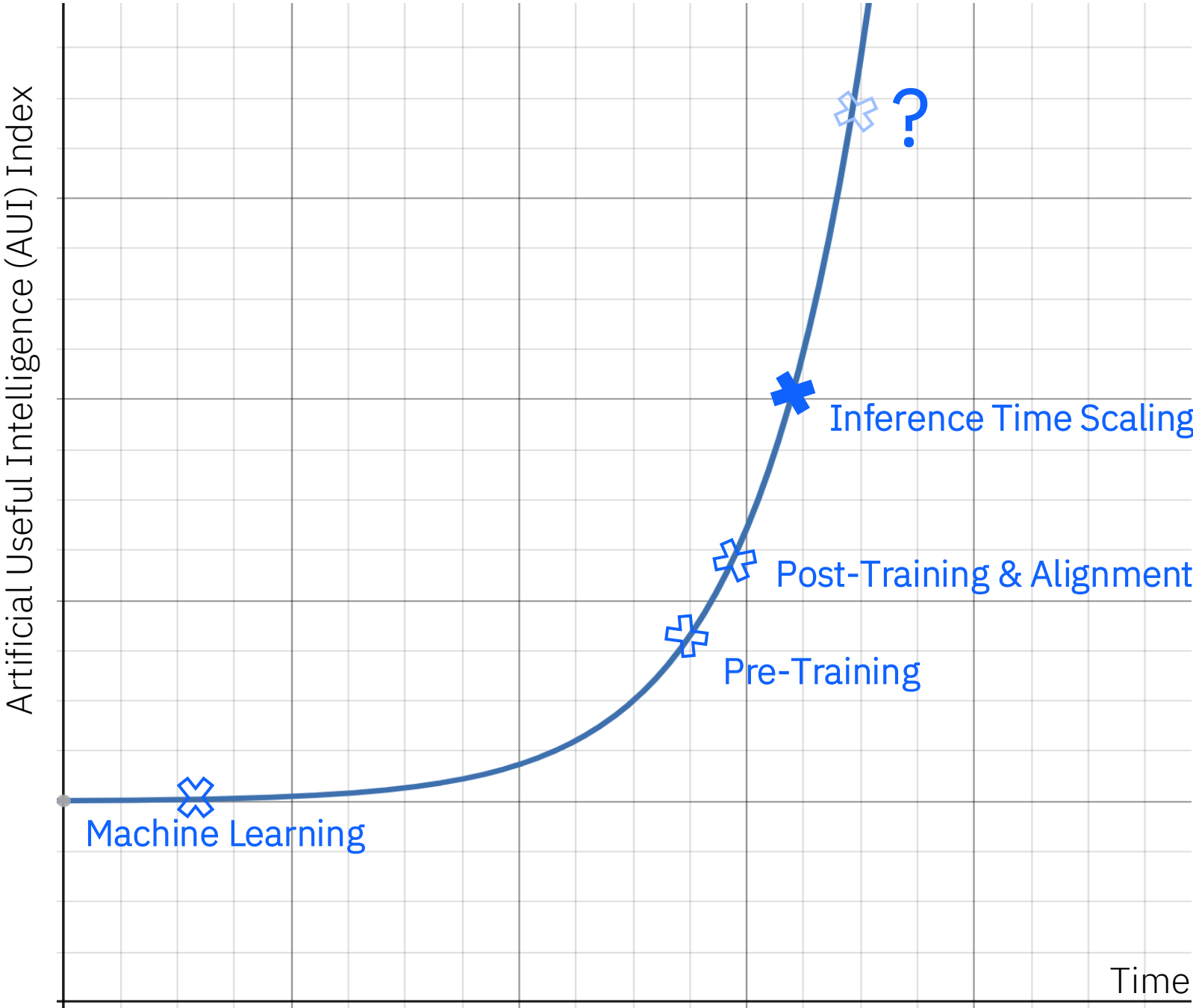
Orchestrator Agent



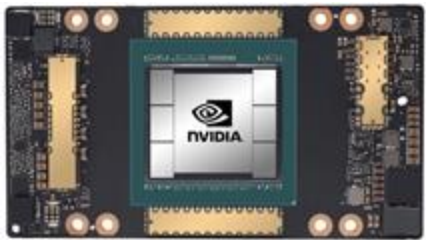
Artificial Intelligence Futures



AI Technology Futures



Hardware Support



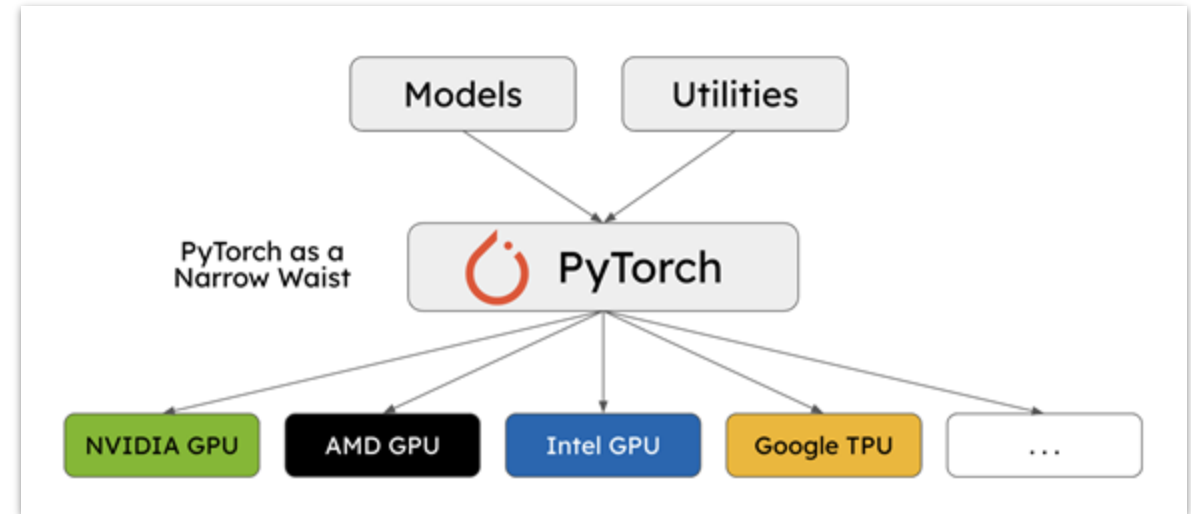
NVIDIA GPU



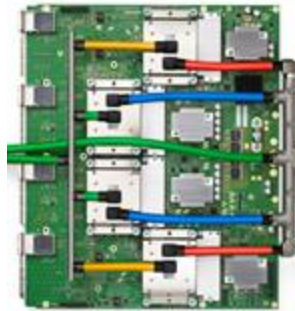
AMD GPU



AWS Neuron



Intel CPU/GPU



Google TPU



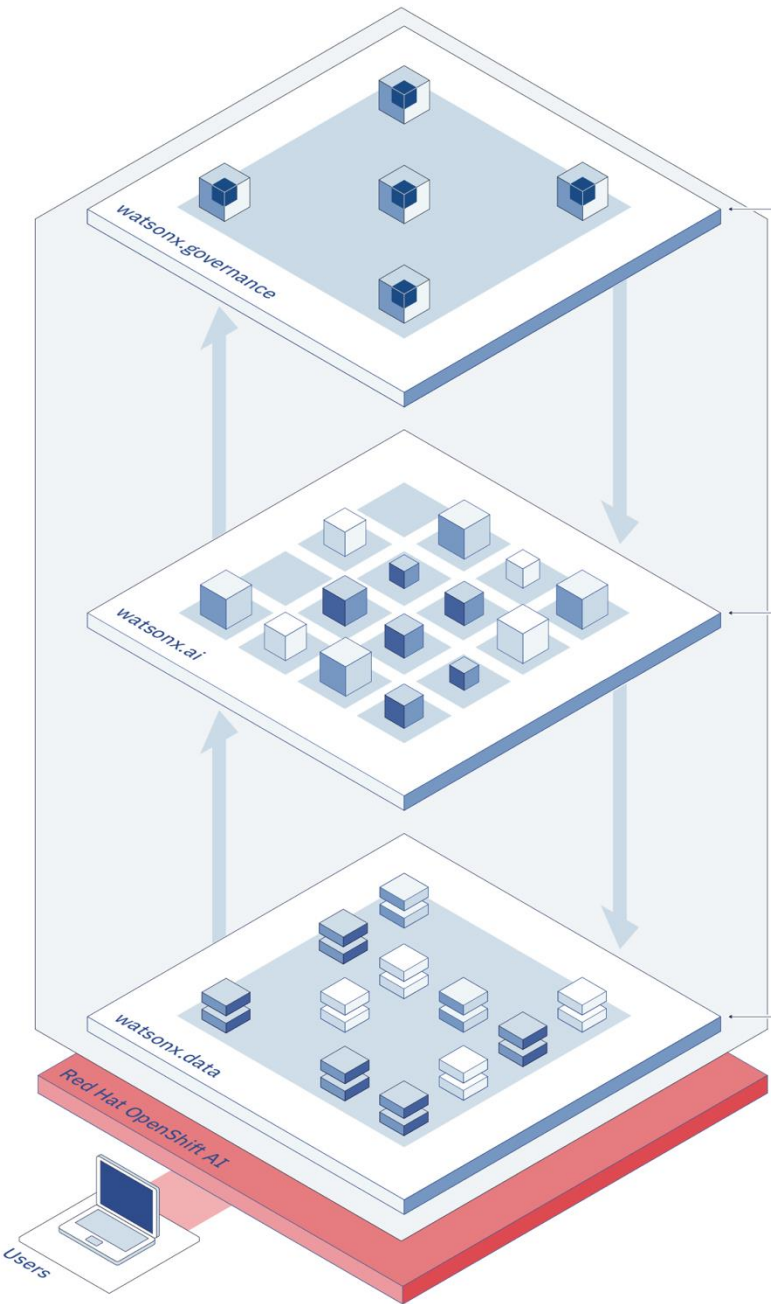
Intel Gaudi



IBM Spyre



watsonx, the platform for AI value creators



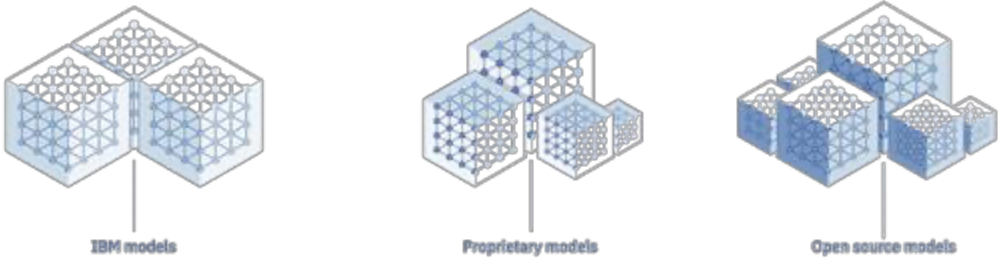
watsonx.governance

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.



Internet		Regulatory	
Code		IT	
		Legal	
Financial		Academic	
		Cyber-security	Climate
Sustainability		News	



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

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.

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-ou0ff/>

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?



mentoria@pullrecast.dev