Preparando-se para o Futuro

AI Leadership Navigator
HORIZONTES DA AI NOS NEGÓCIOS

cnex | MITSloan Management Review Brasil

Alan Braz

Xillenial 1980

Campinas

Sofredor do Guarani

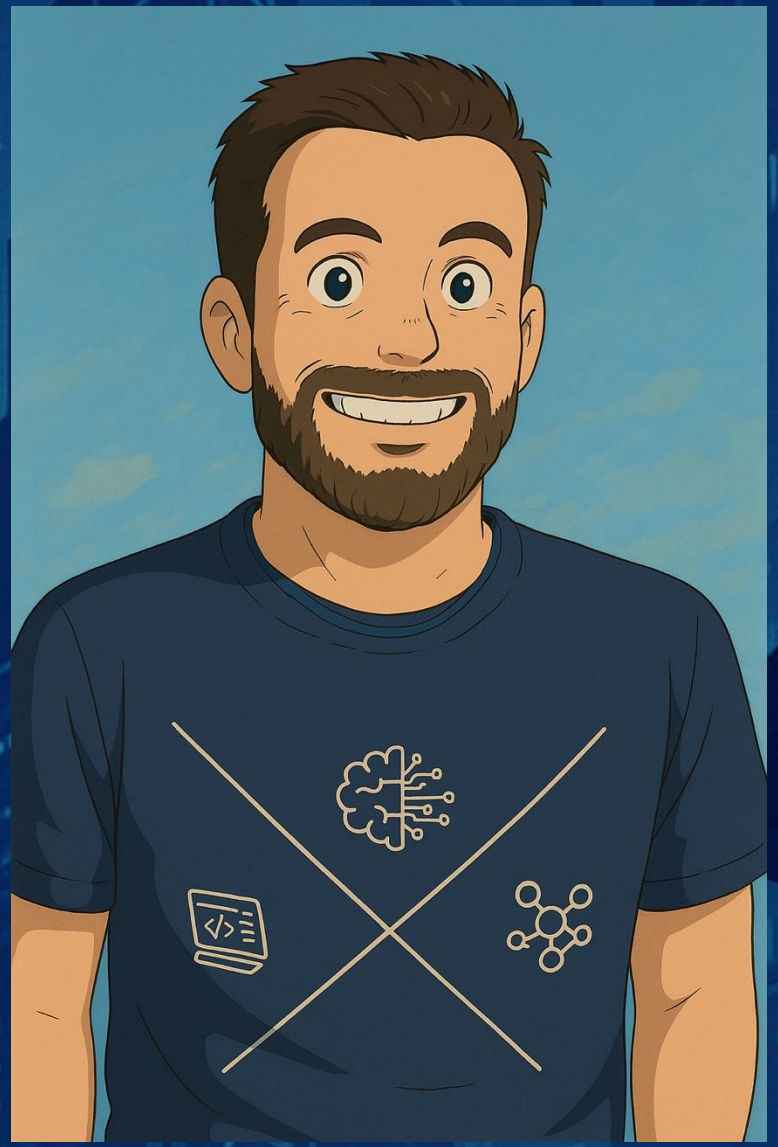
Unicamp - CC MSc

2 filhos – Luiza e Mateus

Esportes NBA/NFL

Fx preta TKD

Empolgado!



alanbraz.com.br / pullrecast.dev

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%

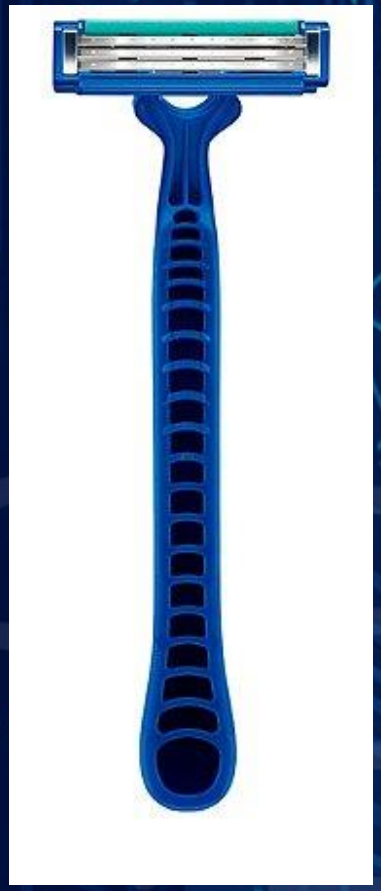
das empresas terão incorporado **Gen AI** em seus processos de negócios em 2026

Fonte: Gartner; PWC report; McKinsey 2022 – 2023; McKinsey 2023 / The Economic Potential of Generative AI


MYTHBUSTERS

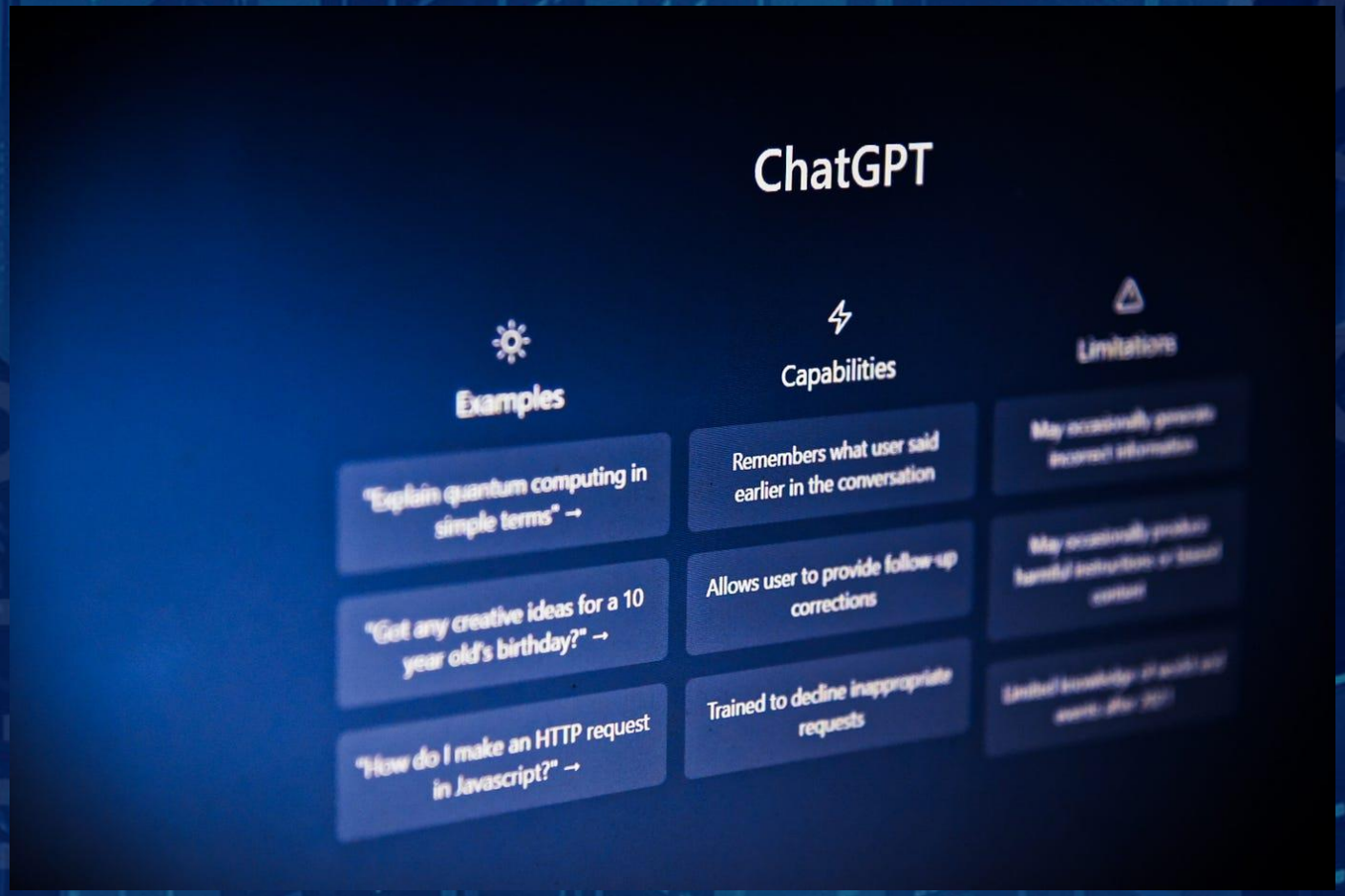

Google
Pesquise ou digite um URL

Sinônimos




ASPIRINA
Yakult
BAND-AID
PLASTIC STRIPS
DANONE
BOM BRIL
CANDIDA
CATUPIRY
SUCRILHOS
MAIZENA


INSTANTÂNEO
Mais Nutritivo
Nestlé
NINHO
Forti+
Fibra
Ferro e Zinco
Vitaminas A, C, D e E
Cálcio
750g
COMPOSTO LÁCTEO COM FIBRAS


ChatGPT
Examples
"Explain quantum computing in simple terms" →
"Got 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

perplexity

Ask anything or @mention a Space

Home
Discover
Spaces
Account
Upgrade

Troubleshoot
Travel

Web
Search across the entire Internet

Academic
Search academic papers

Social
Discussions and opinions

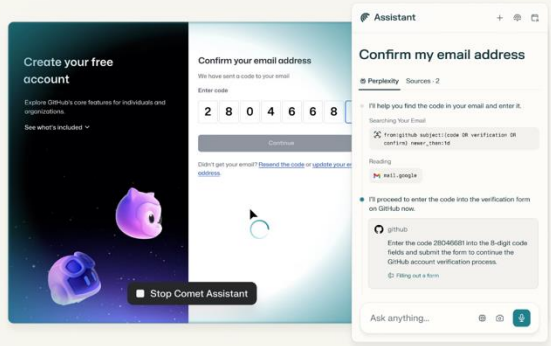
Finance
Search SEC filings

Comet is

powerful

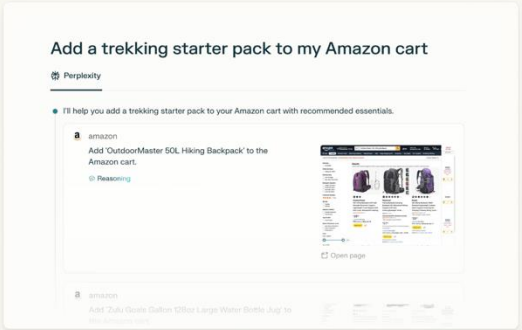
Unified AI search, instant context, and automation across every site. Summarize, shop, schedule, and research—directly in the browser.

Download



Enjoy the web and let the AI assistant handle your chores

It clicks, types, submits and autofills — so you don't have to.



Do everything on your to-do list faster with Comet

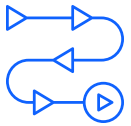
Ask Comet to handle complex tasks like shopping, from comparing products and reading reviews, through to checkout.

https://www.perplexity.ai/

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

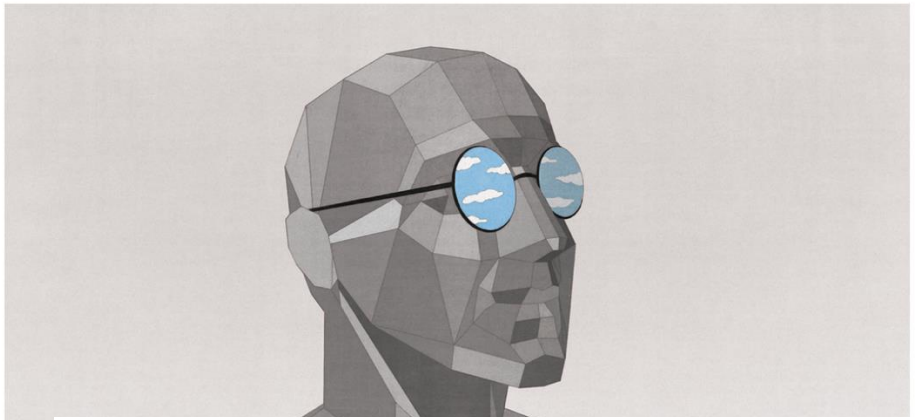
Criação de código

Generative AI

How People Are Really Using Gen AI in 2025

by Marc Zao-Sanders

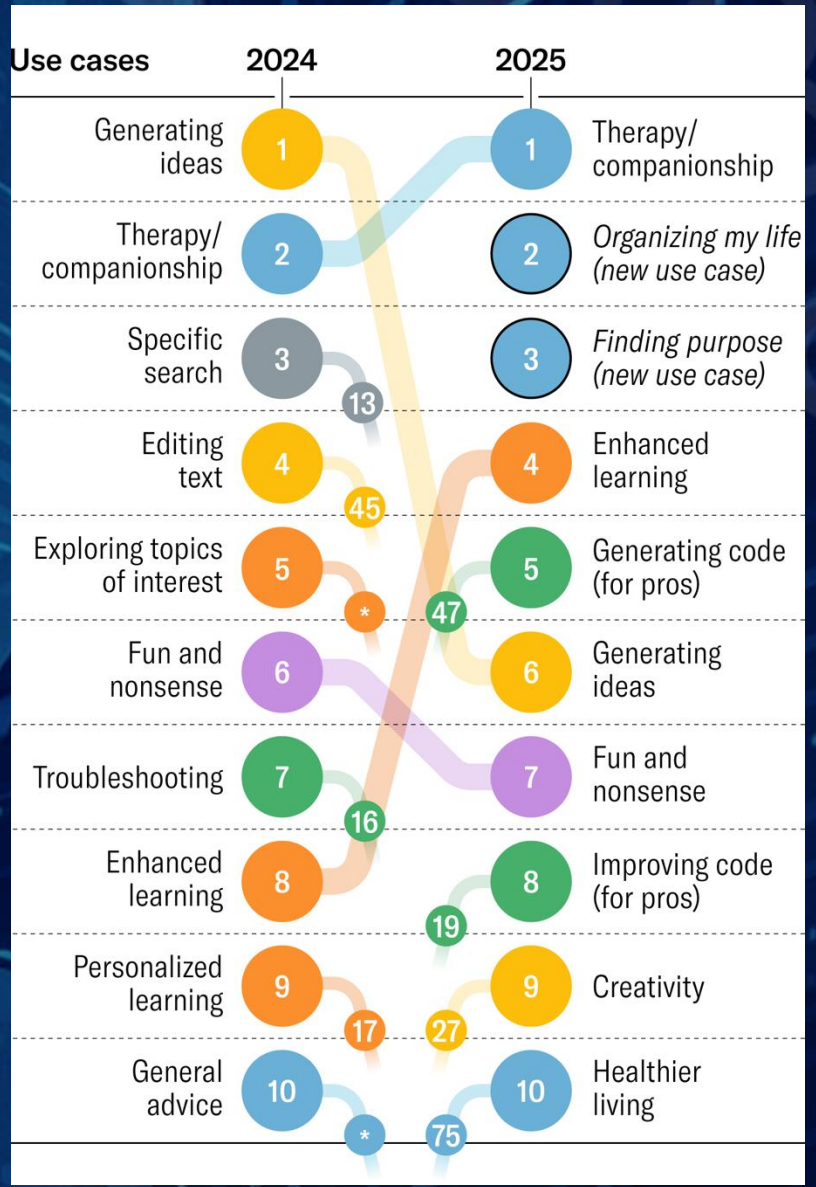
April 9, 2025



Illustrat

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.



https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

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 VINY MATHIAS | ATUALIZADO: 25 DE MAIO DE 2025 ÀS 16:01

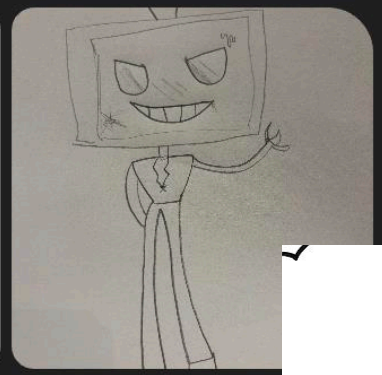
https://br.ign.com/tech/140588/news/69-dos-membros-da-geracao-z-pedem-coisas-ao-chatgpt-educadamente-com-uma-intencao-oculta

ChatGPT 4o

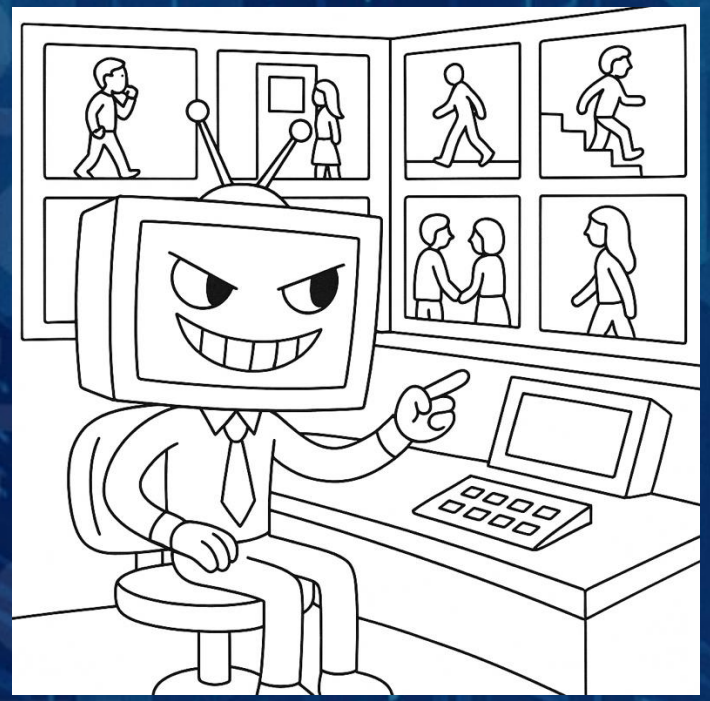
Share

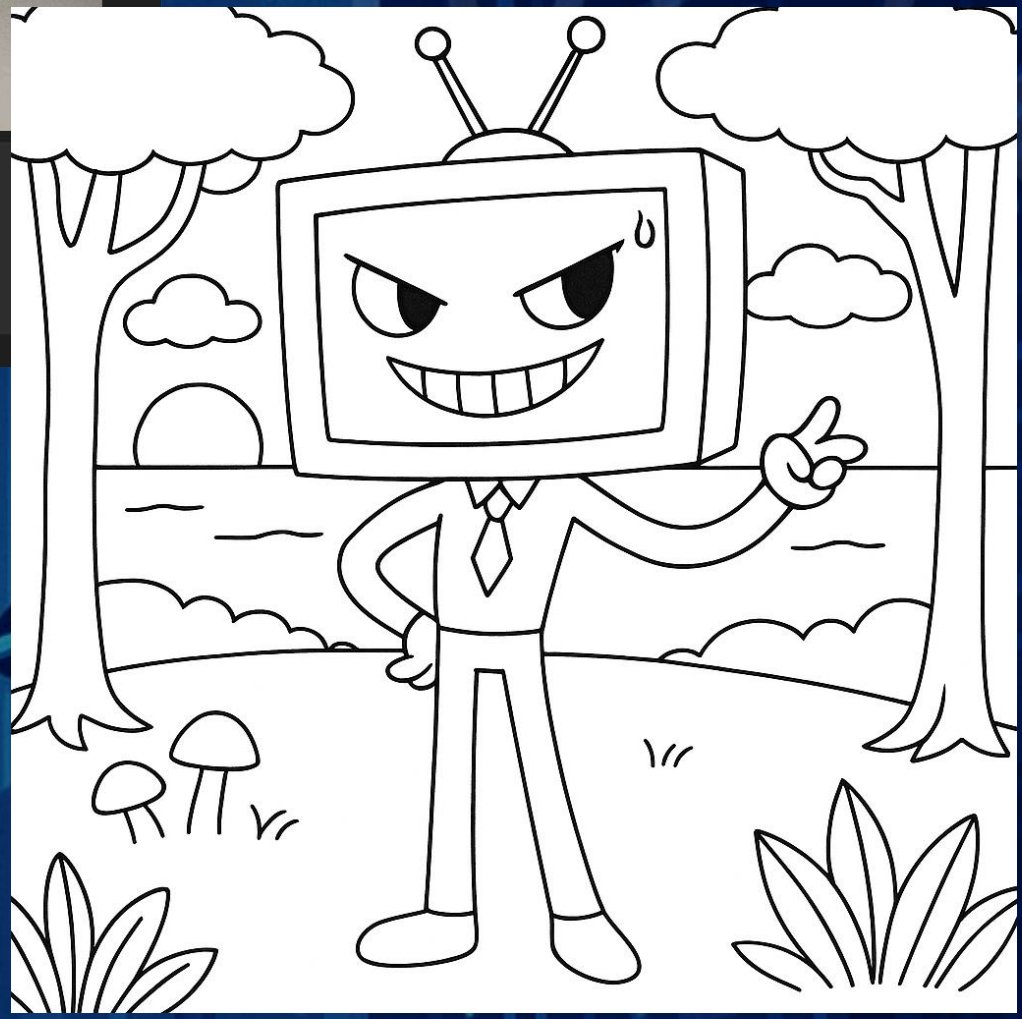
PLUS

create a scene with these 2 characters fighting with swords



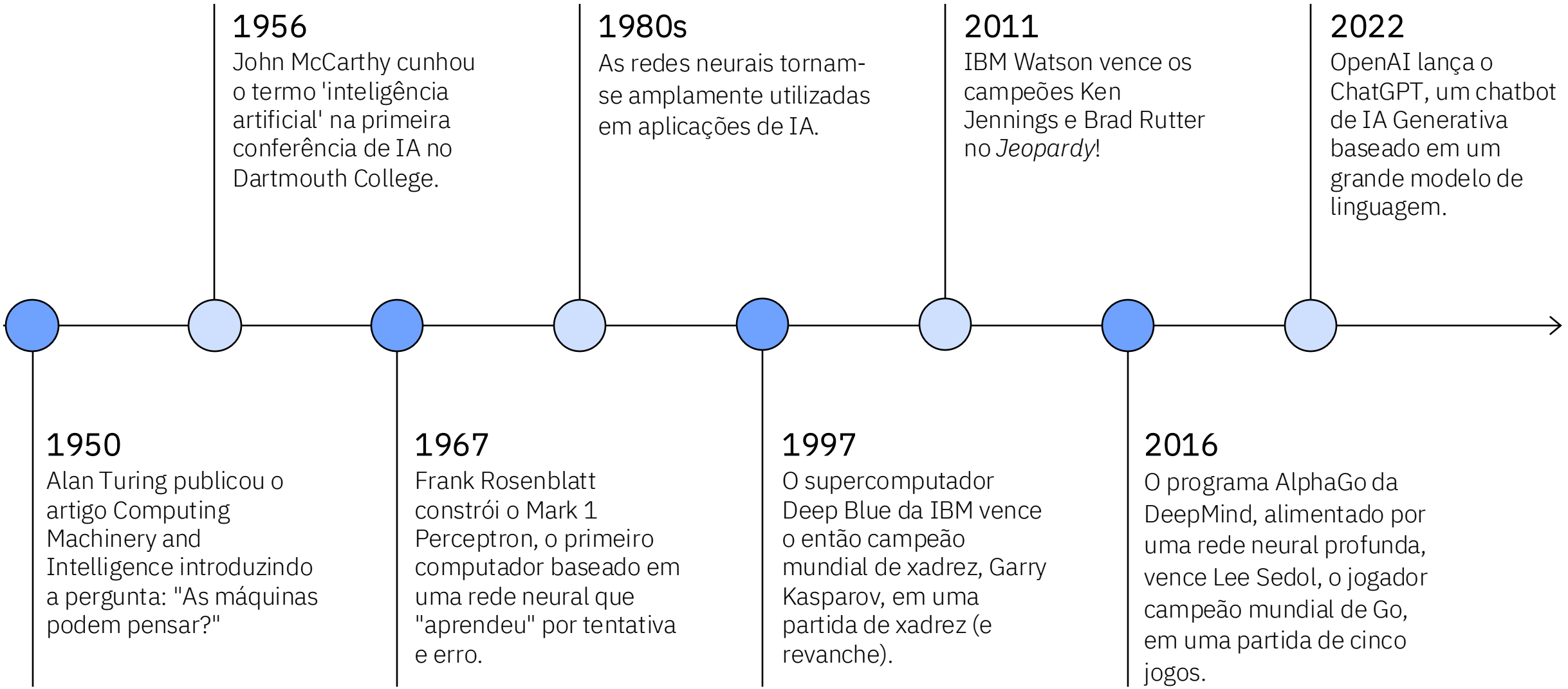
Use the cute bears coloring page as reference of style and a coloring page scene using the mad tv caracter as the main theme





alanbraz.com.br / pullrecast.dev

Marcos da IA


1950
Alan Turing publicou o artigo Computing Machinery and Intelligence introduzindo a pergunta: "As máquinas podem pensar?"
1956
John McCarthy cunhou o termo 'inteligência artificial' na primeira conferência de IA no Dartmouth College.
1967
Frank Rosenblatt constrói o Mark 1 Perceptron, o primeiro computador baseado em uma rede neural que "aprendeu" por tentativa e erro.
1980s
As redes neurais tornam-se amplamente utilizadas em aplicações de IA.
1997
O supercomputador Deep Blue da IBM vence o então campeão mundial de xadrez, Garry Kasparov, em uma partida de xadrez (e revanche).
2011
IBM Watson vence os campeões Ken Jennings e Brad Rutter no *Jeopardy*!
2016
O programa AlphaGo da DeepMind, alimentado por uma rede neural profunda, vence Lee Sedol, o jogador campeão mundial de Go, em uma partida de cinco jogos.
2022
OpenAI lança o ChatGPT, um chatbot de IA Generativa baseado em um grande modelo de linguagem.

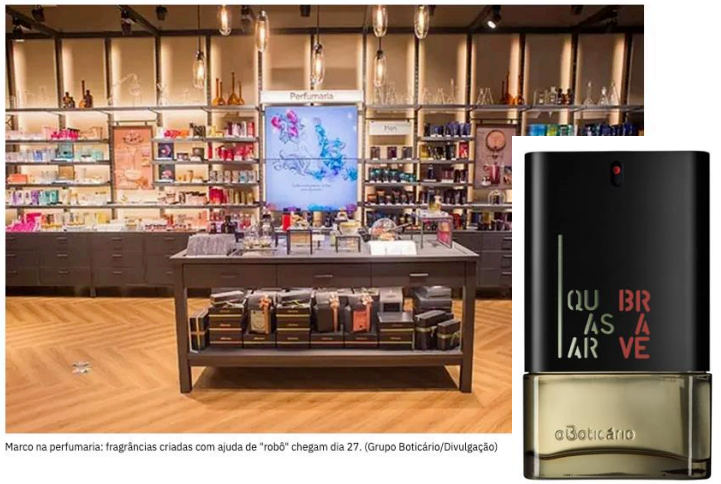
Deep Blue 1997

Ei! 2013/2014

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

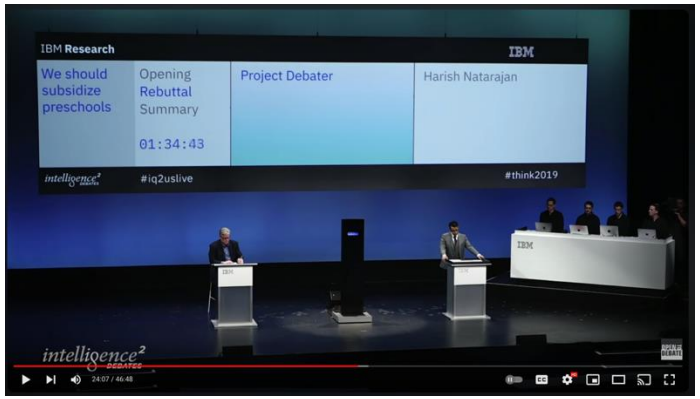


Marco na perfumaria: fragrâncias criadas com ajuda de "robô" chegam dia 27. (Grupo Boticário/Divulgação)

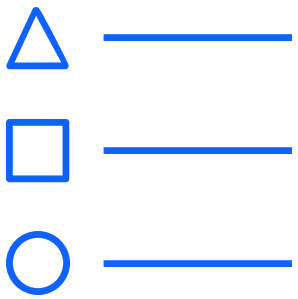
Watson 2011

Chef Watson 2017

Project Debater



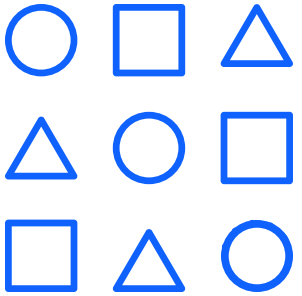
Sistemas baseados em regras



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

1950's 1980's 2010's 2020's 2025's

82%

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

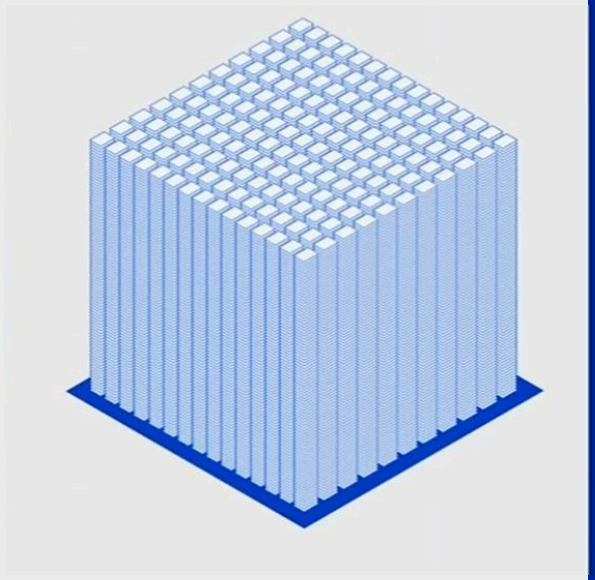
55%

Líderes empresariais sem informações importantes sobre seus gastos com tecnologia

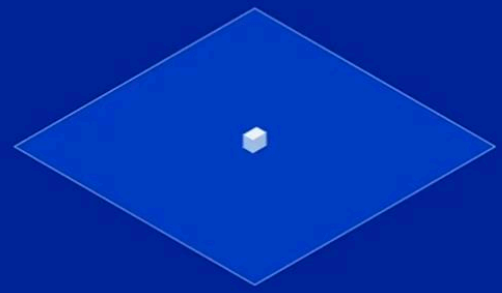
10%

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



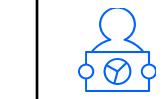
Sources: IBM Institute for Business Value; Gartner 2023

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 facilmanete os resultados e decisões indicados pela IA.

Justa

Treinada em dados imparciais.

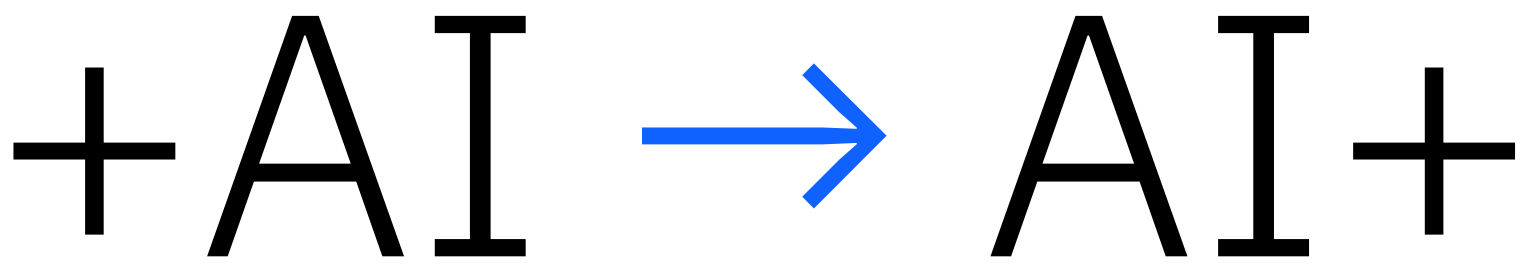
Robusta

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

Privada

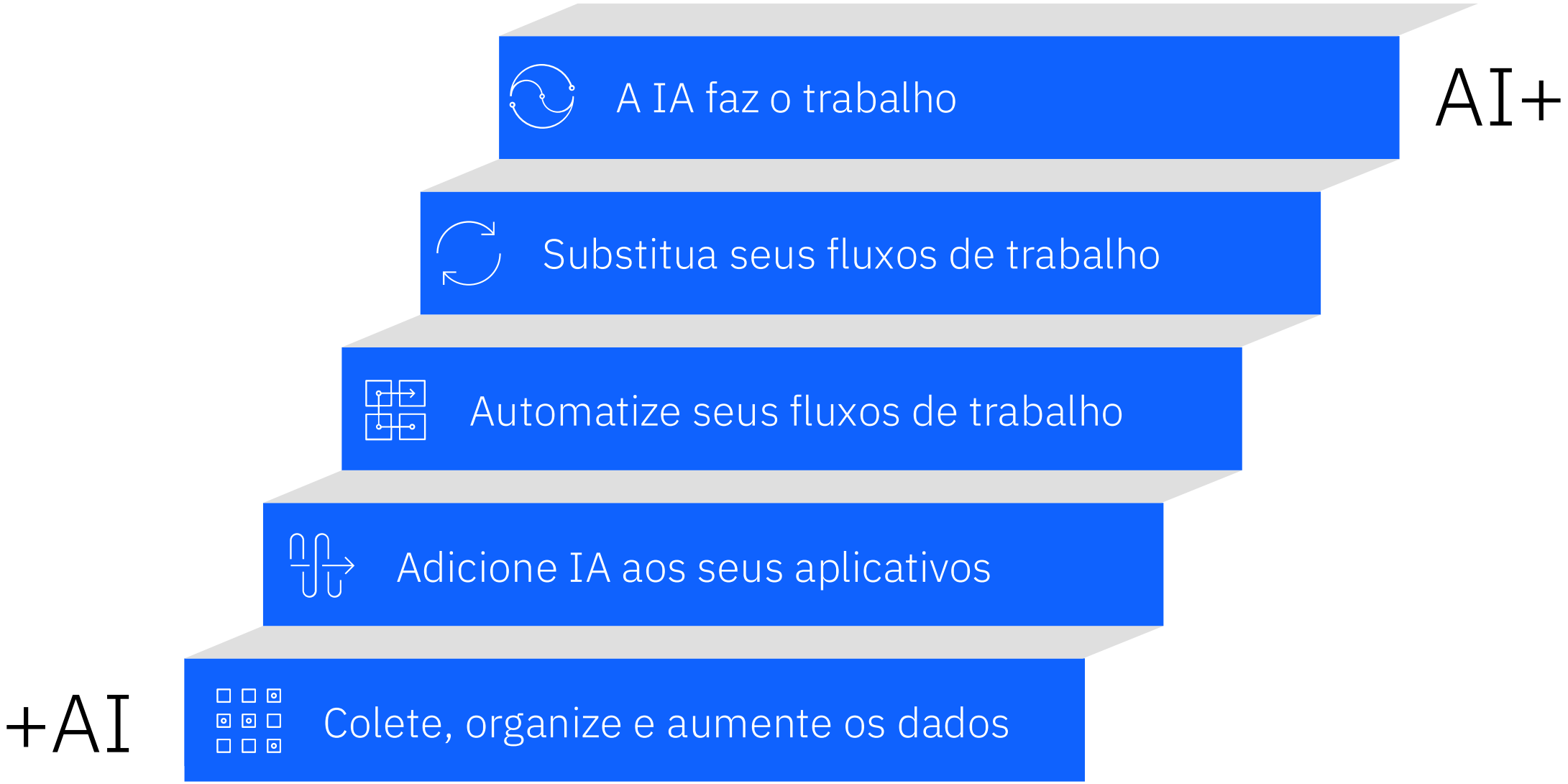
Alimentada por dados de alta integridade compatíveis com os negócios e defende a privacidade e necessidades de segurança.

EFICIENTE



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

A escada da IA moderna



Employee Experience ↷

AskHR

Um único canal: experiência envolvente

2 camadas: Digital & Humano

IBM

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

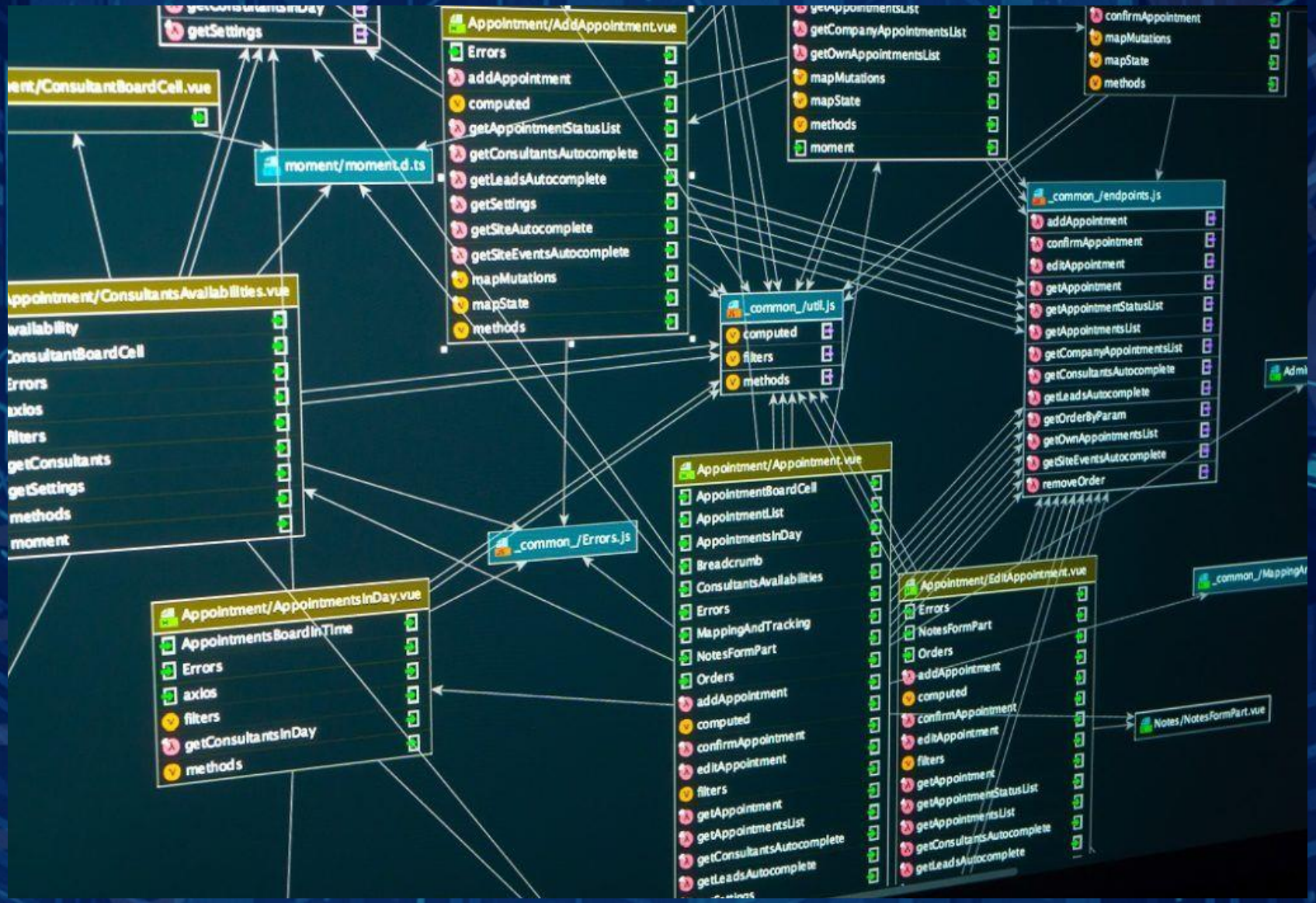
75% Melhora em produtividade

61% Redução de abertura de chamados

75% Feedbacks positivos (CSAT)

89% de interações resolvidas diretamente via AskHR

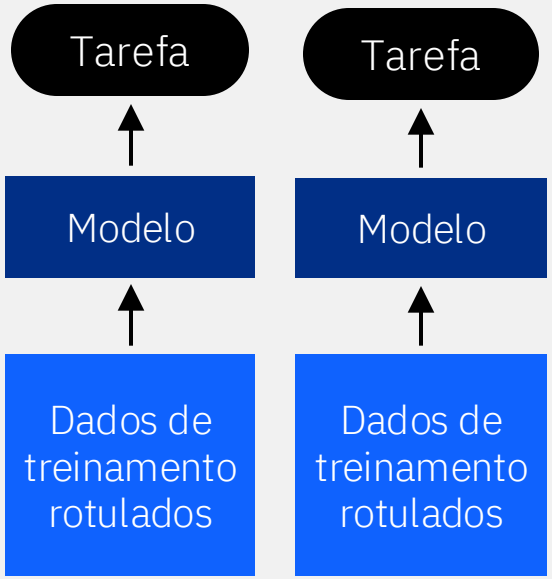
Dados de 2023


getSettings
Appointment/AddAppointment.vue
Errors
addAppointment
computed
getAppointmentStatusList
getConsultantsAutocomplete
getLeadsAutocomplete
getSettings
getSiteAutocomplete
getSiteEventsAutocomplete
mapMutations
mapState
methods
moment/moment.d.ts
getCompanyAppointmentsList
getOwnAppointmentsList
mapMutations
mapState
methods
moment
confirmAppointment
mapMutations
mapState
methods
common/endpoints.js
addAppointment
confirmAppointment
editAppointment
getAppointment
getAppointmentStatusList
getAppointmentsList
getCompanyAppointmentsList
getConsultantsAutocomplete
getLeadsAutocomplete
getOrderByParam
getOwnAppointmentsList
getSiteEventsAutocomplete
removeOrder
common/util.js
computed
filters
methods
Appointment/ConsultantsAvailabilities.vue
availability
ConsultantBoardCell
Errors
axios
filters
getConsultants
getSettings
methods
moment
common/Errors.js
Appointment/Appointment.vue
AppointmentBoardCell
AppointmentList
AppointmentsInDay
Breadcrumb
ConsultantsAvailabilities
Errors
MappingAndTracking
NotesFormPart
Orders
addAppointment
computed
confirmAppointment
editAppointment
filters
getAppointment
getAppointmentsList
getConsultantsAutocomplete
getLeadsAutocomplete
Appointment/AppointmentsInDay.vue
AppointmentsBoardInTime
Errors
axios
filters
getConsultantsInDay
methods
Appointment/EditAppointment.vue
Errors
NotesFormPart
Orders
addAppointment
computed
confirmAppointment
editAppointment
filters
getAppointment
getAppointmentStatusList
getAppointmentsList
getConsultantsAutocomplete
getLeadsAutocomplete
Notes/NotesFormPart.vue

O que vem a seguir com os modelos de fundação:

IA convencional

1000s – 100000s Pontos de dados rotulados / Tarefa



Zero-shot prompting

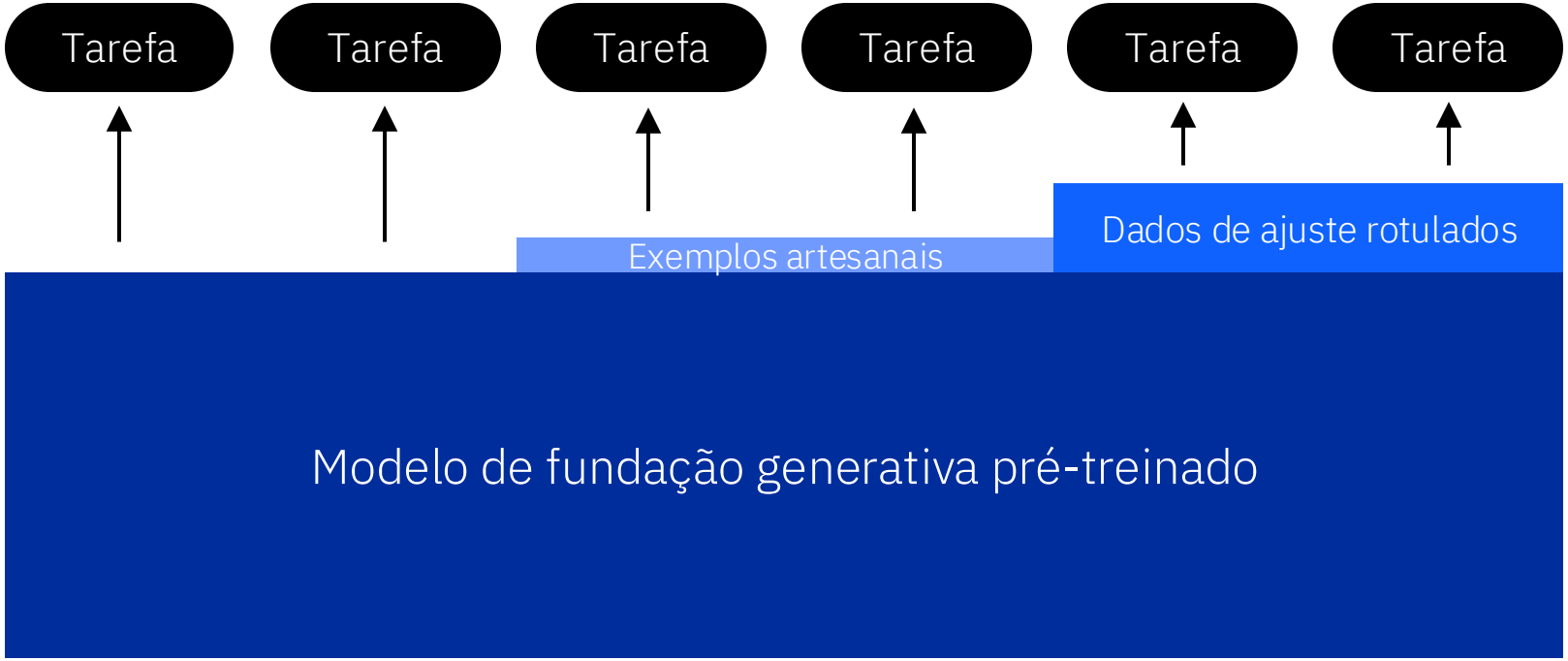
No labeled data needed

Few-shot prompting

1-10 labeled data points / task

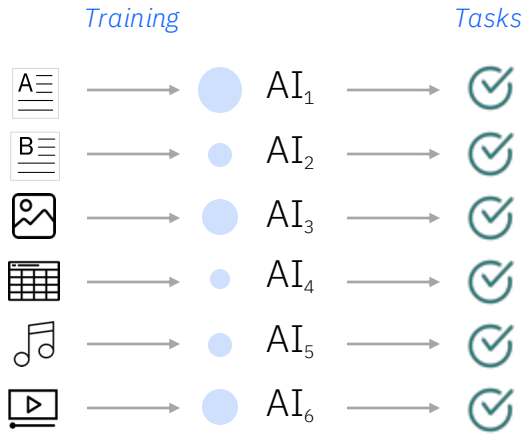
Prompt-tuning

100s – 1000s labeled data points / task



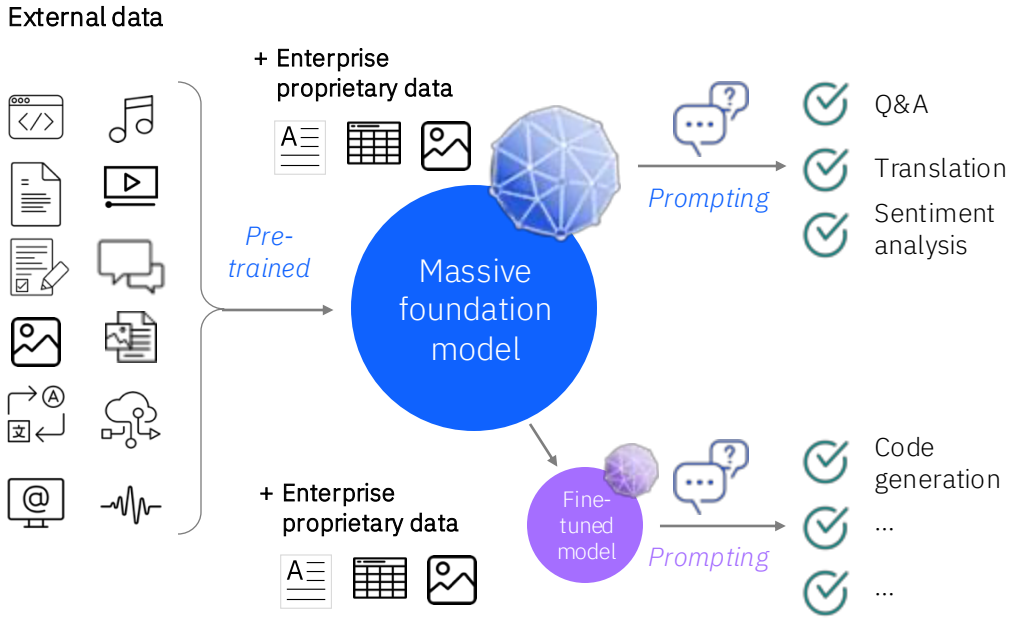
Foundation models establish a new paradigm for AI capabilities

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

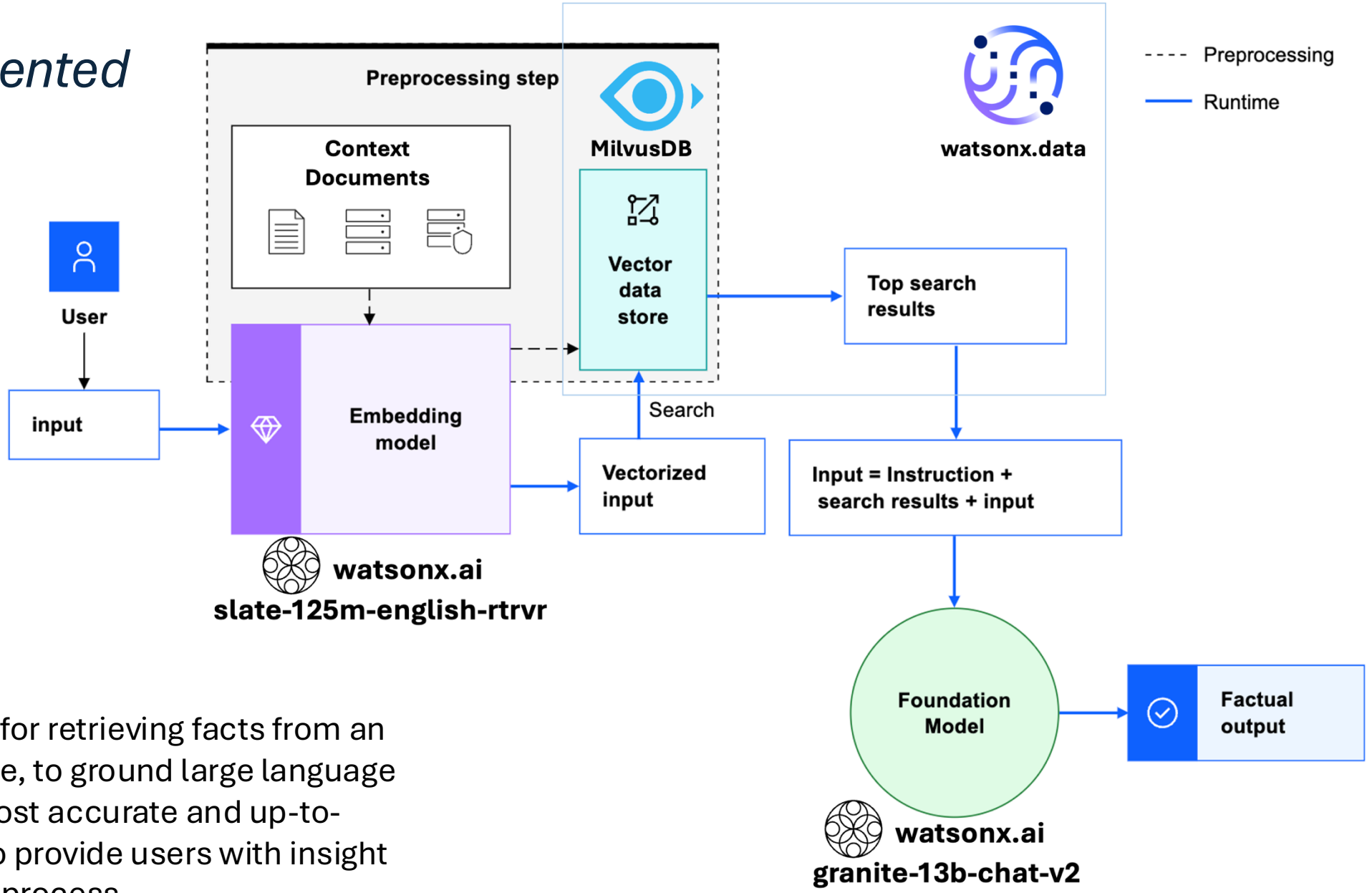
- Summarization
- Conversational knowledge
- Content creation
- Code co-creation

Key advantages

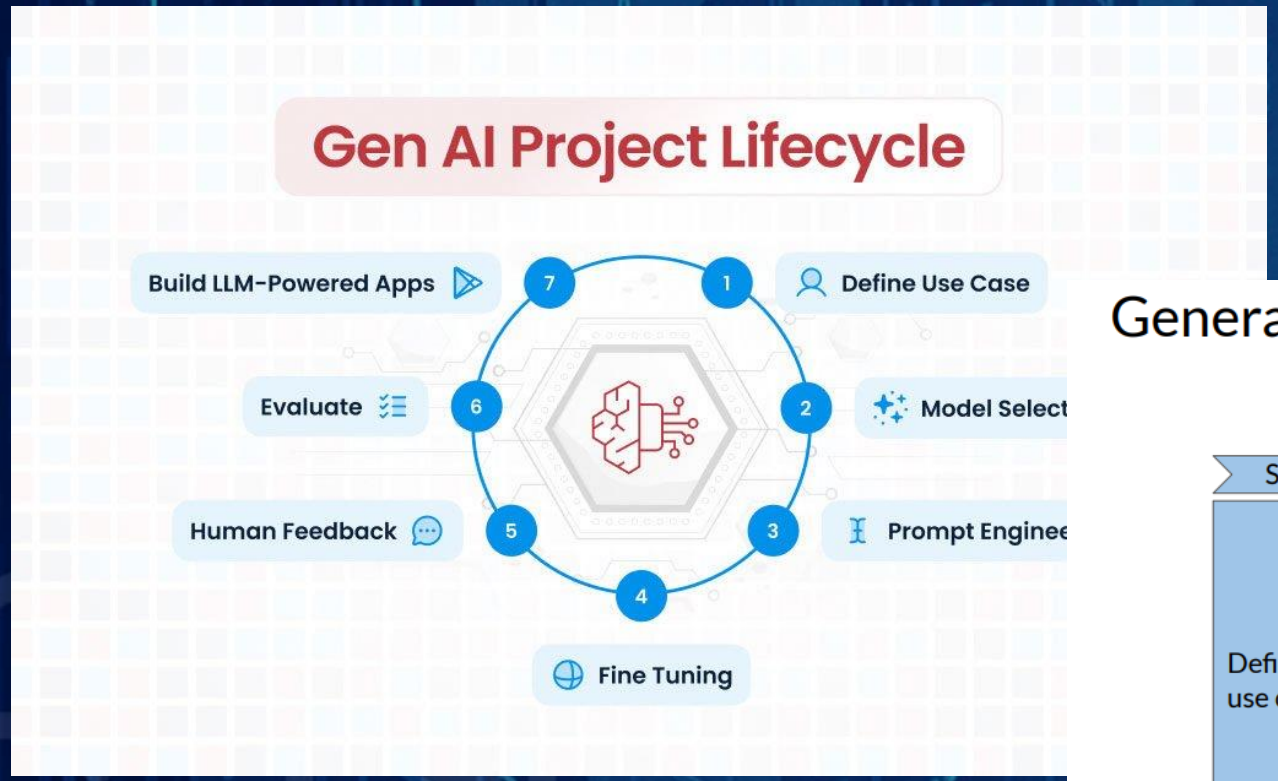
- Lower upfront costs through less labeling
- Faster deployment through fine tuning
- Equal or better accuracy for multiple use cases
- Incremental revenue through better performance

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.



Generative AI project lifecycle

Scope	Select	Adapt and align model		Application integration	
Define the use case	Choose an existing model or pretrain your own	Prompt engineering Fine-tuning Align with human feedback	Evaluate	Optimize and deploy model for inference	Augment model and build LLM-powered applications

DeepLearning.AI aws

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

Quantum computing

Hybrid cloud

Security

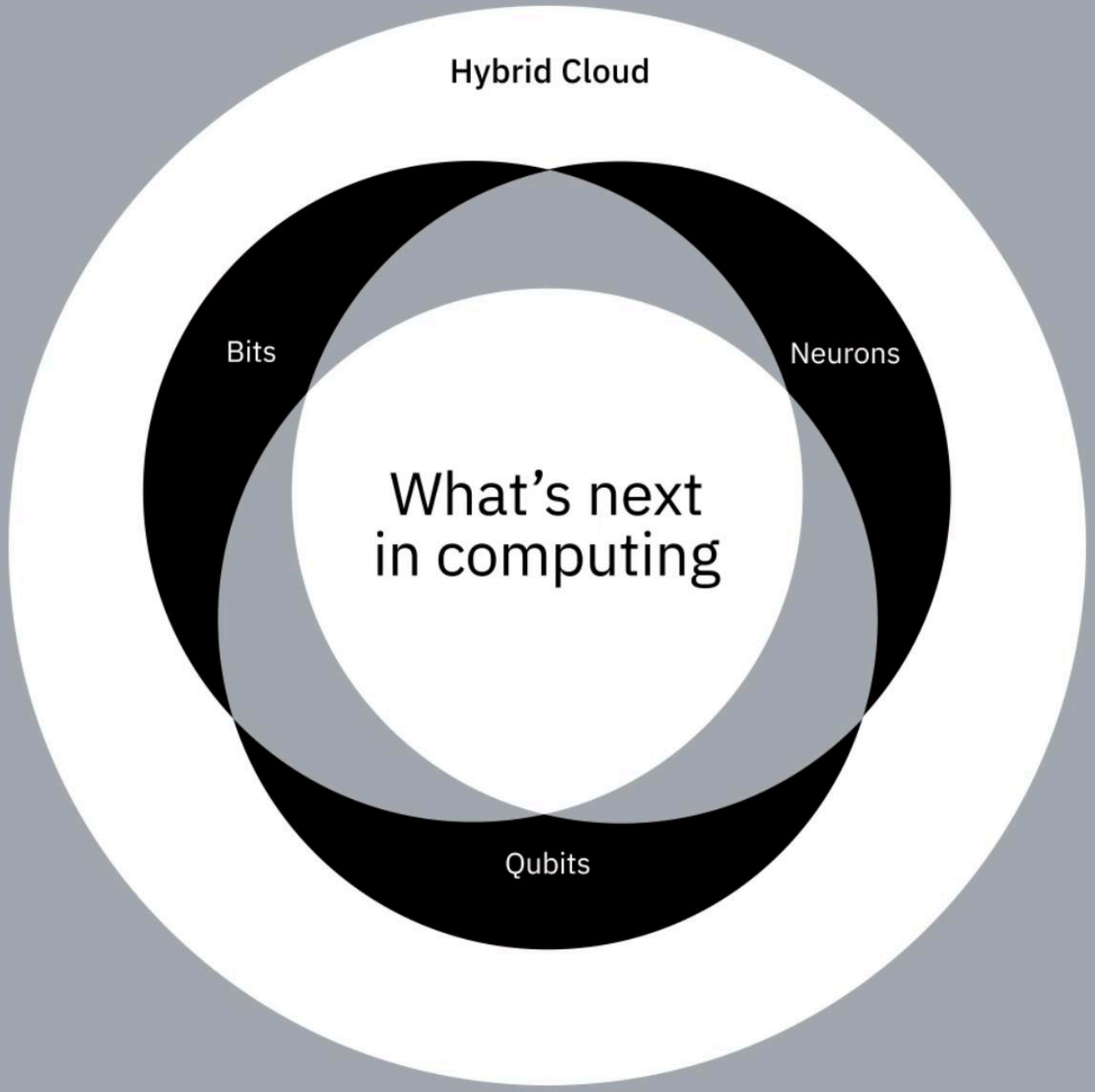
Africa
Albany
Almaden
Brazil
Cambridge
India
Ireland
Israel
Tokyo
United Kingdom
Yorktown Heights
Zurich

6 Nobel Laureates

5 National Medals of Science

10 Medals of Technology

6 Turing Awards

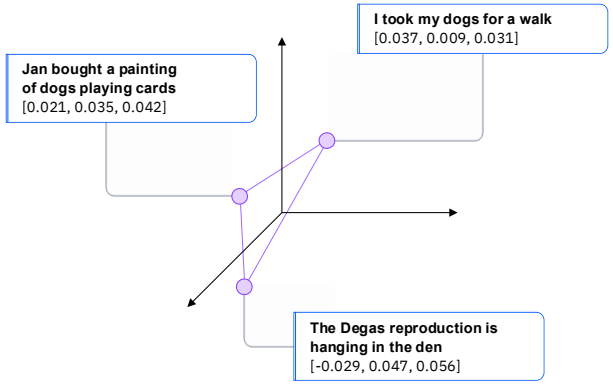


~~watsonx~~ AI Platform Engineering

- 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

vLLM

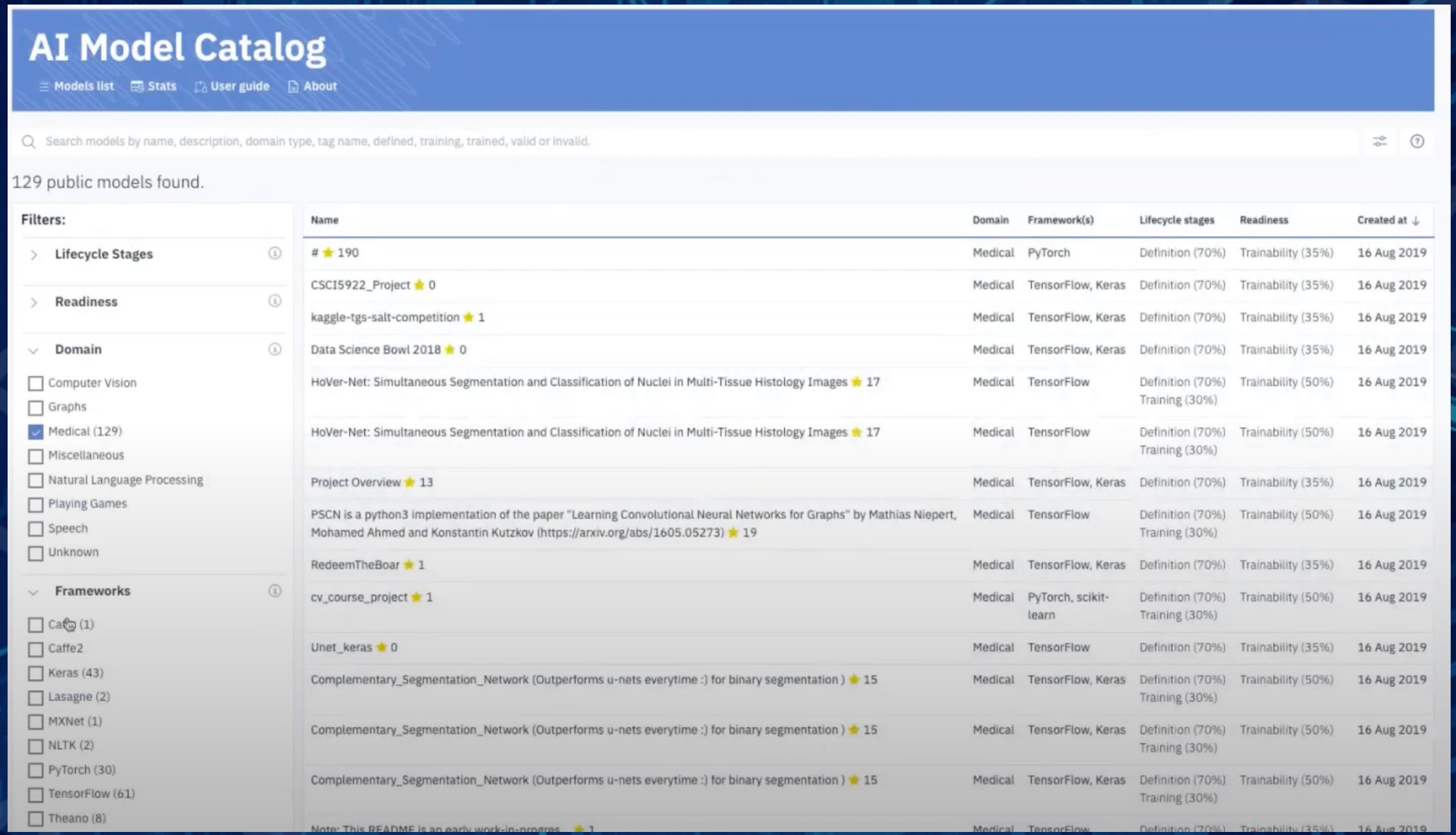
* OpenSource

Compiler/Driver

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

Rodando e otimizando para AIU/Spyre

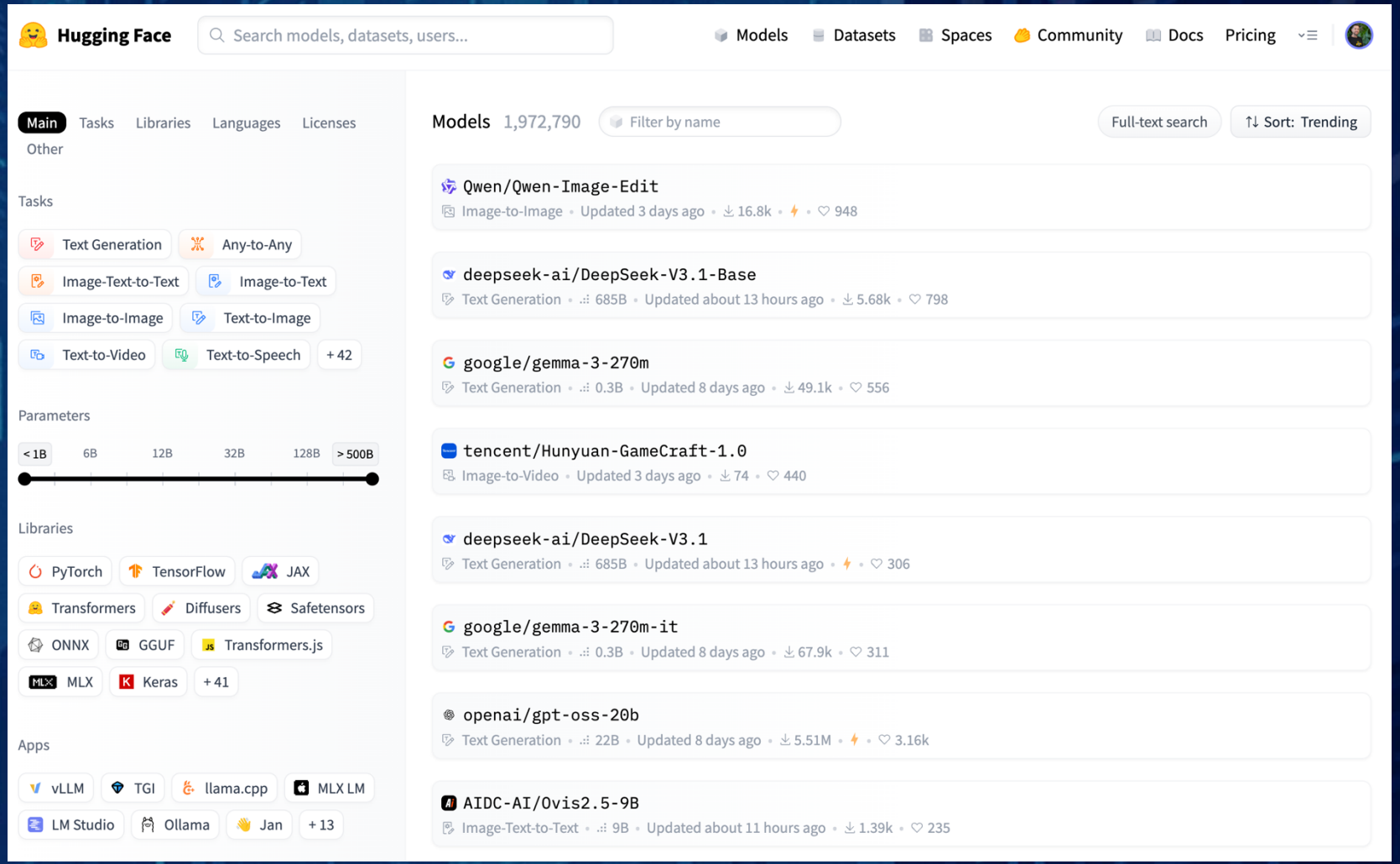
2017


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%) 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

Hugging Face

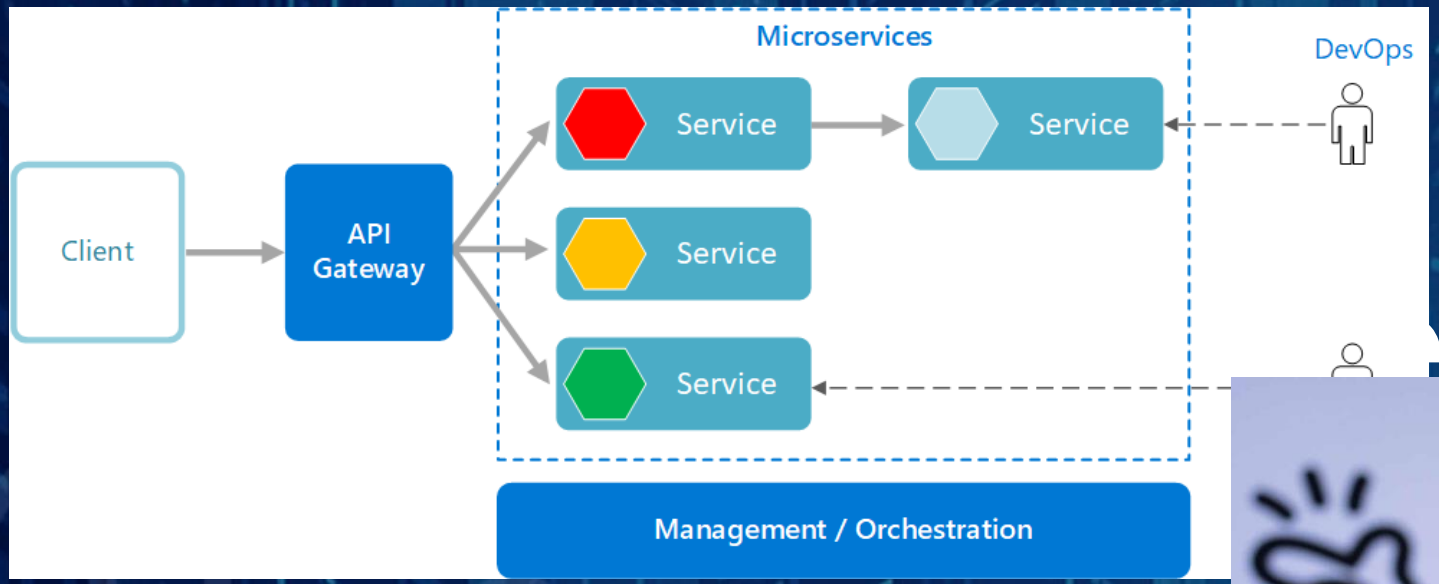
Models 1,972,790

Datasets 480,099



https://huggingface.co/

21 de Agosto de 2025


Microservices
DevOps
Client
API Gateway
Service
Service
Service
Service
Management / Orchestration

TA


manus
The general AI agent

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

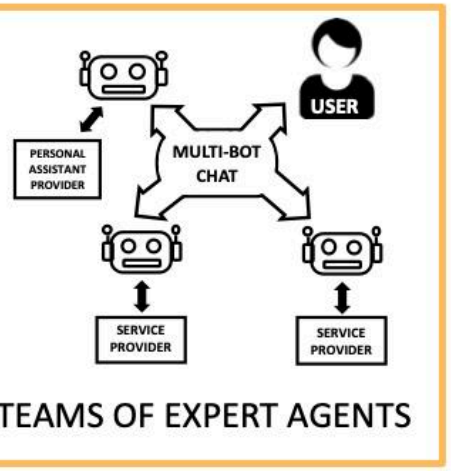
- Criptografia pós-quântica
- Inteligência ambiental invisível
- Computação verde
- Computação híbrida

Sinergia entre humanos e máquinas

- Computação espacial
- Robôs polifuncionais
- Aprimoramento neurológico

Fonte: Gartner

Gartner®

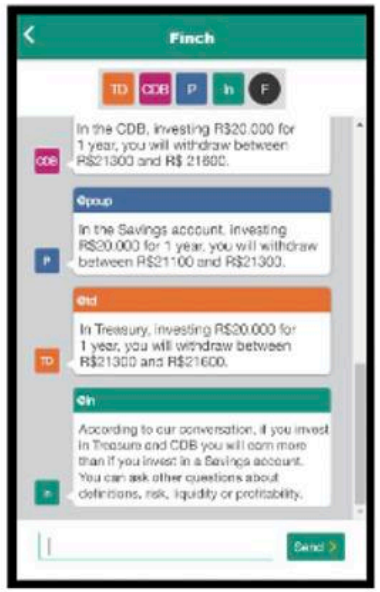


TEAMS OF EXPERT AGENTS

cdbGuru
reactive bot

poupancaGuru
reactive bot

tesouroGuru
reactive bot



from Q&A dialog to multiparty chat

User

investmentGuru
proactive bot

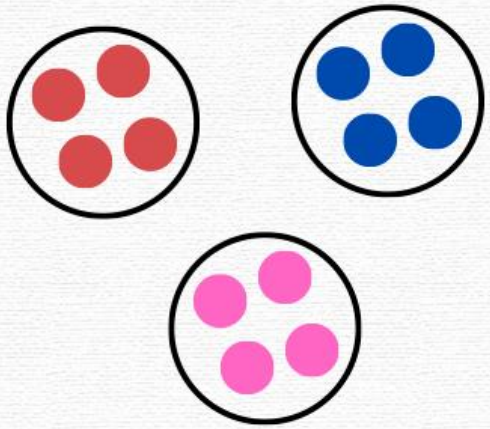
15

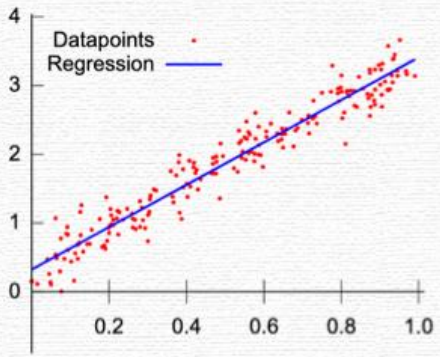
Self-service finance advising using multiple chatbots

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

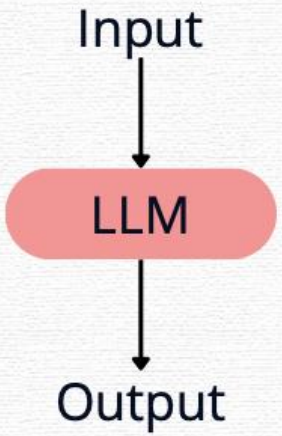
Evolução da Inteligência Artificial

Classificação/Predição

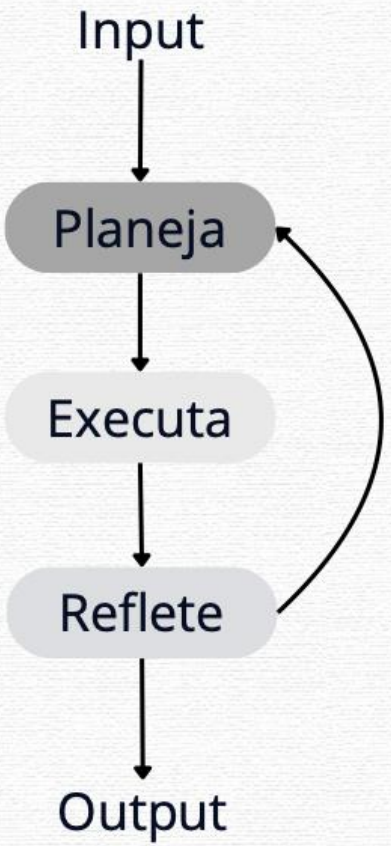




Assistants



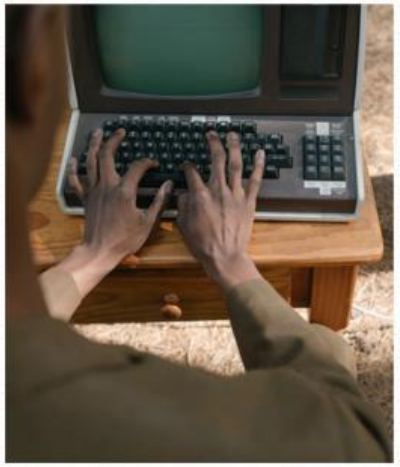
AI Agent

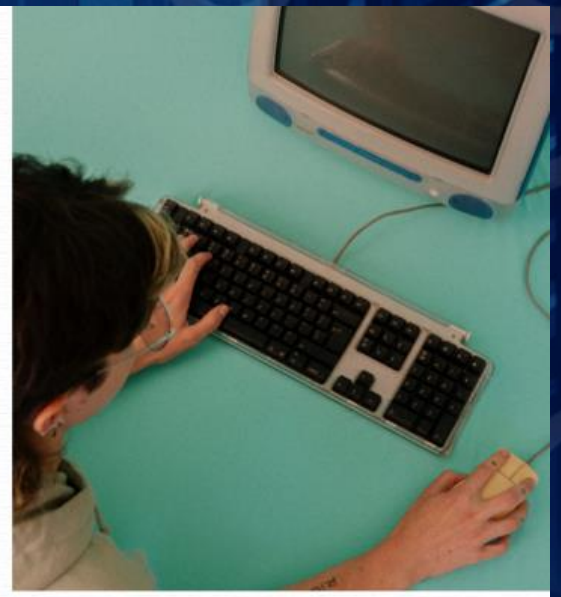


Introdução

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.

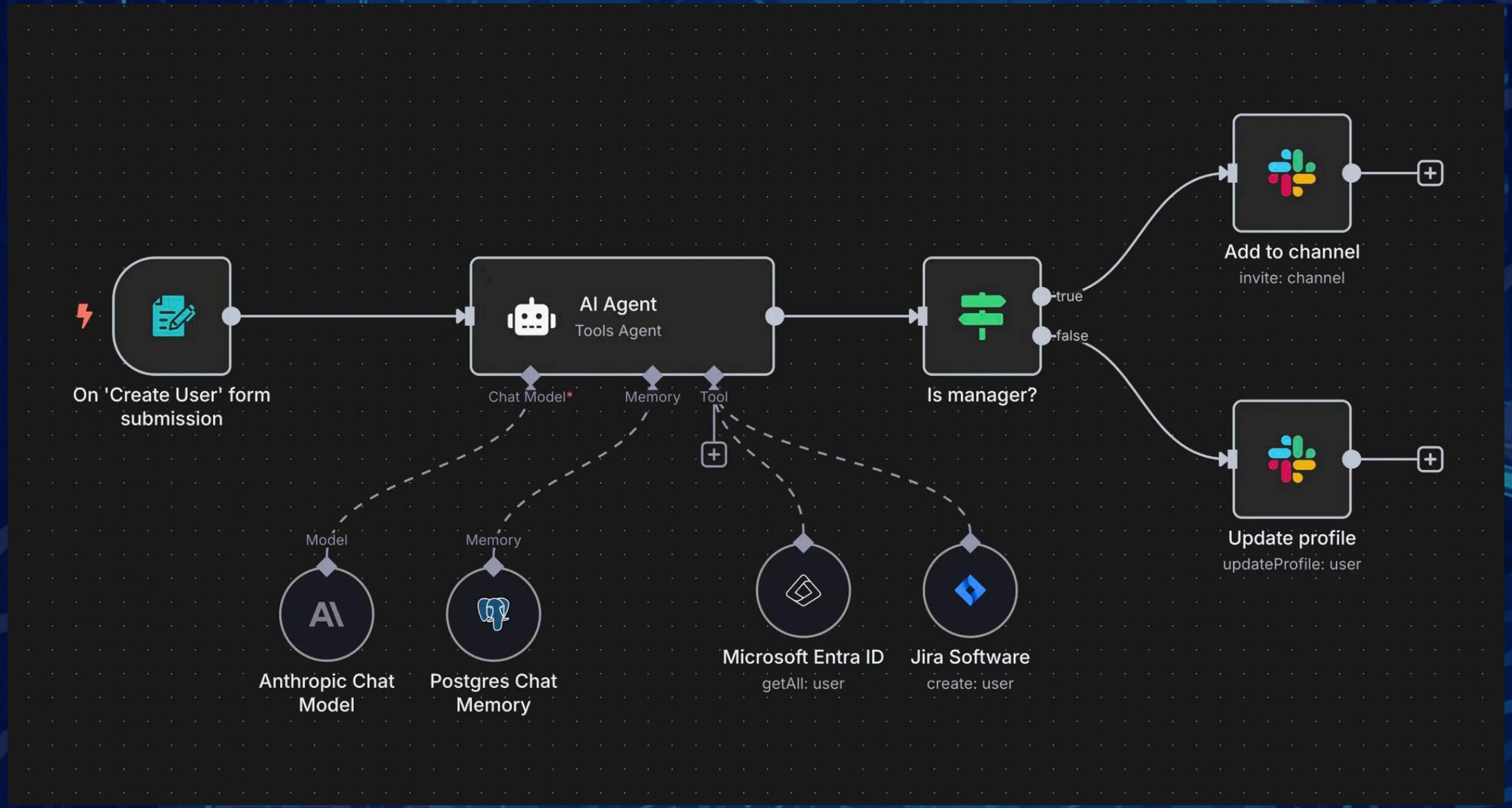




Quick demo - Full visibility of the interim steps

```
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": "celsiu
> Agent: (tool_output) {"latitude":36.16438,"longitude":-115.14392,"generationtime_ms":0.08392333984375,"utc_offset_seconds":0,"timezone":"UTC","timezone_abbrev
al":"seconds","temperature_2m":"∞C","rain":"mm","apparent_temperature":"∞C"},"current":{"time":"2024-09-13T11:15","interval":900,"temperature_2m":22.3,"rain":0,
ure_2m":"∞C","relative_humidity_2m":"%","apparent_temperature":"∞C"},"hourly":{"time":["2024-09-13T00:00","2024-09-13T01:00","2024-09-13T02:00","2024-09-13T03:0
T07:00","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",
: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,
midity_2m":[9,9,10,10,11,12,13,13,11,11,11,11,11,13,15,14,11,12,11,10,10,9,8,8],"apparent_temperature":[30.1,30.2,27.9,26.4,25.7,25.3,23.2,22.3,20.7,19.8,19.7,1
units":{"time":"iso8601","apparent_temperature_max":"∞C","apparent_temperature_min":"∞C","sunrise":"iso8601","sunset":"iso8601"},"daily":{"time":["2024-09-13"],
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.
```

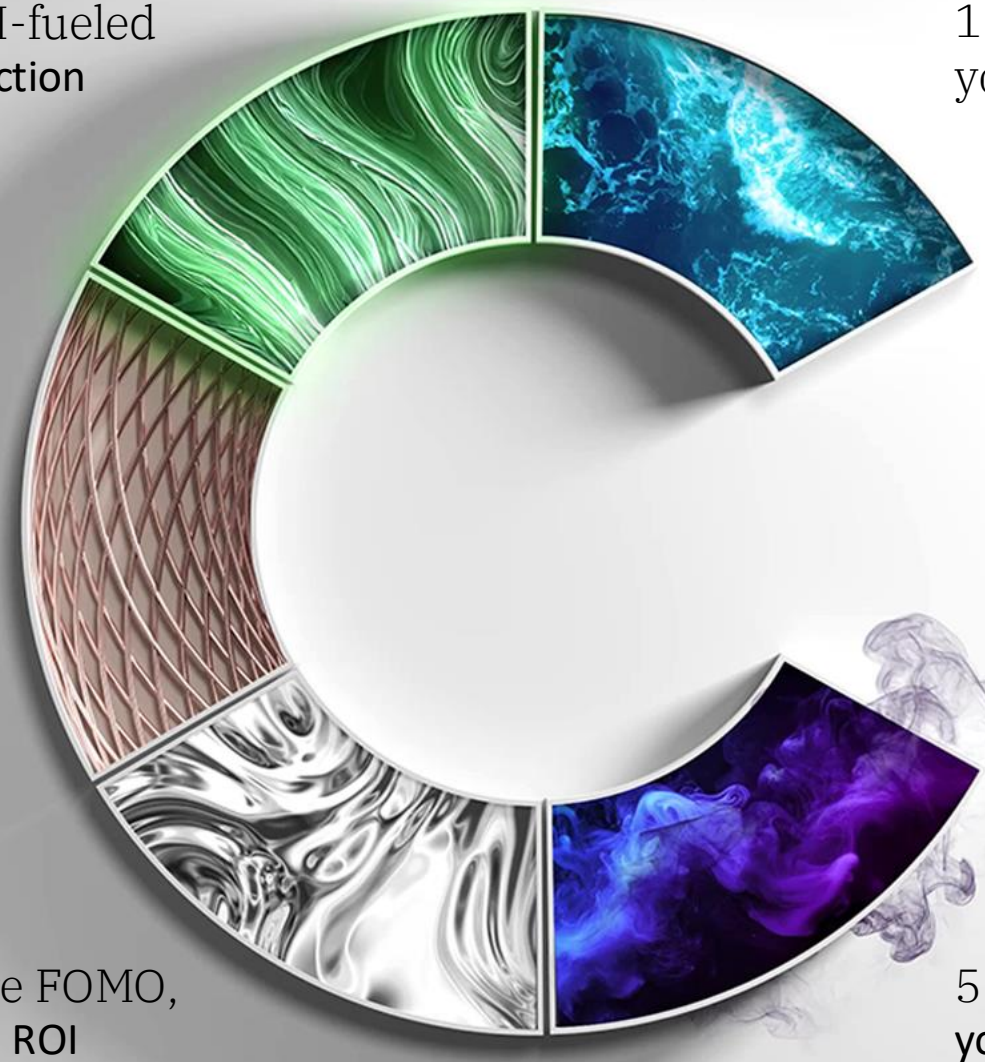

On 'Create User' form submission
AI Agent
Tools Agent
Chat Model*
Memory
Tool
Is manager?
true
false
Add to channel
invite: channel
Update profile
updateProfile: user
Model
Memory
Anthropic Chat Model
Postgres Chat Memory
Microsoft Entra ID
getAll: user
Jira Software
create: user

https://n8n.io/

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**

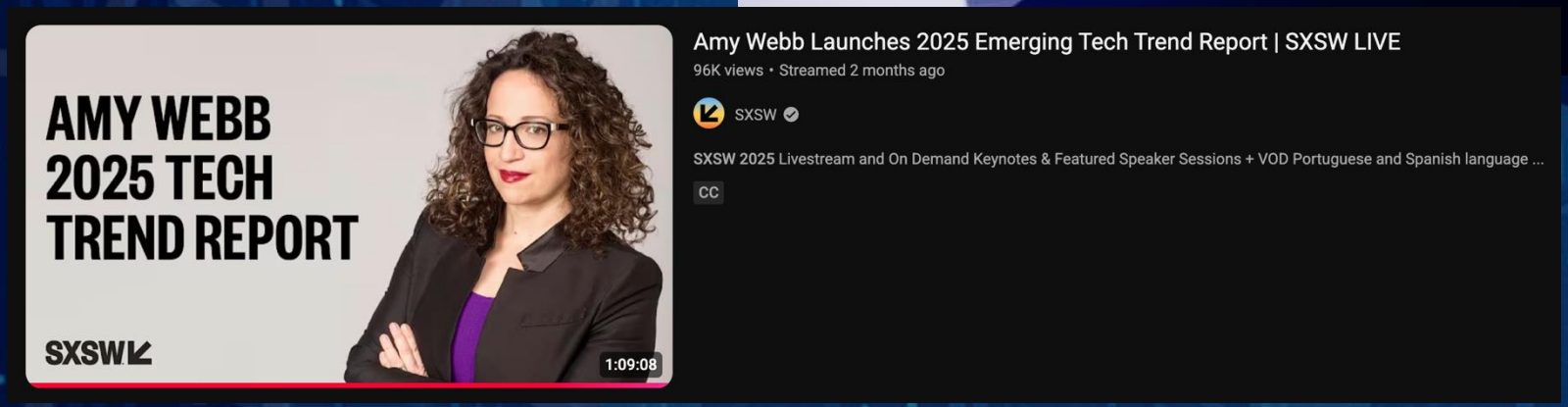
4. Ignore FOMO, **lean into ROI**

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



https://ftsg.com/wp-content/uploads/2025/03/FTSG_2025_TR_FINAL_LINKED.pdf

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.**

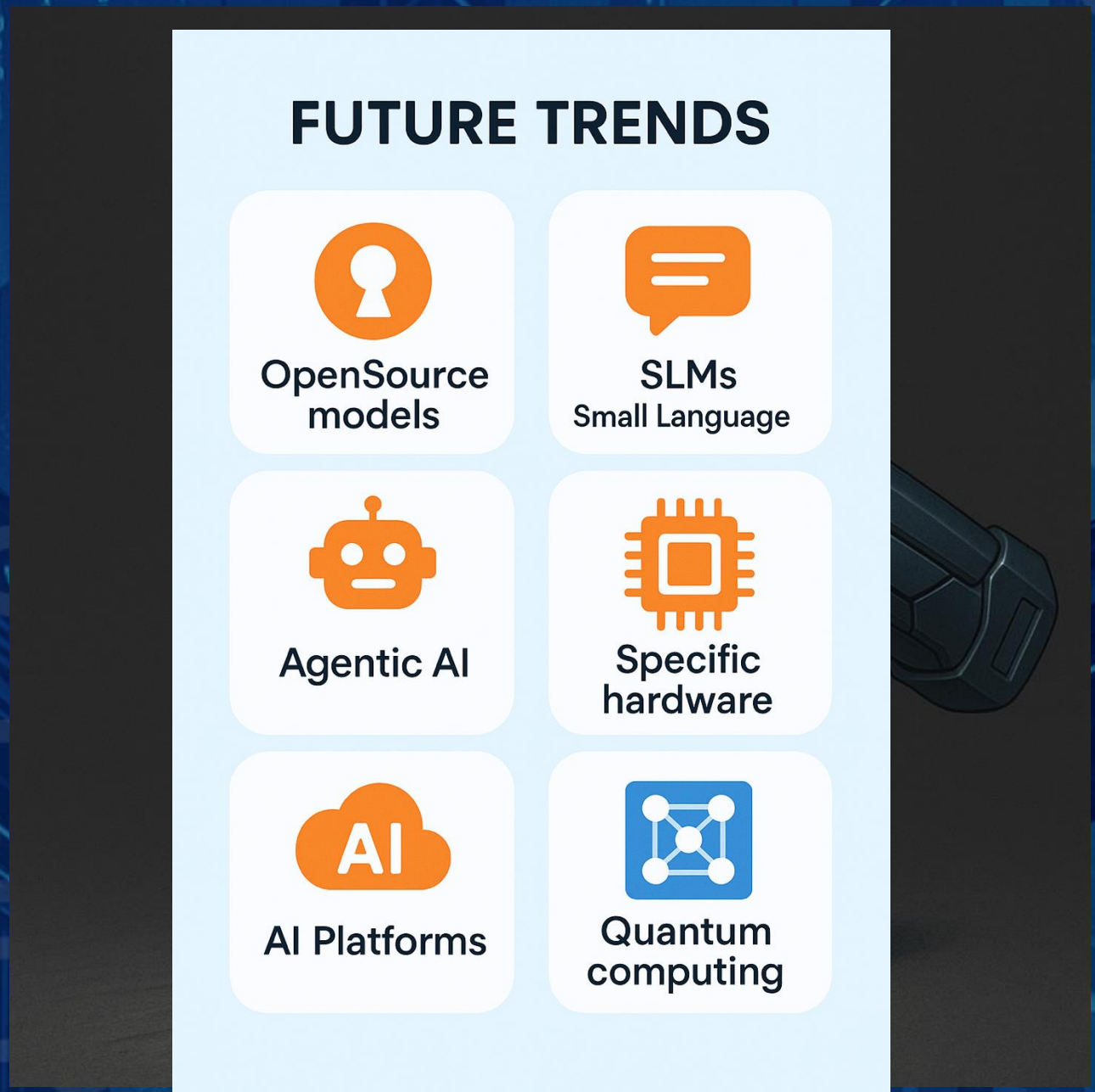
10 Key Takeaways from the FTSG 2025 Tech Trends Report..

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-IyLW

alanbraz.com.br / pullrecast.dev

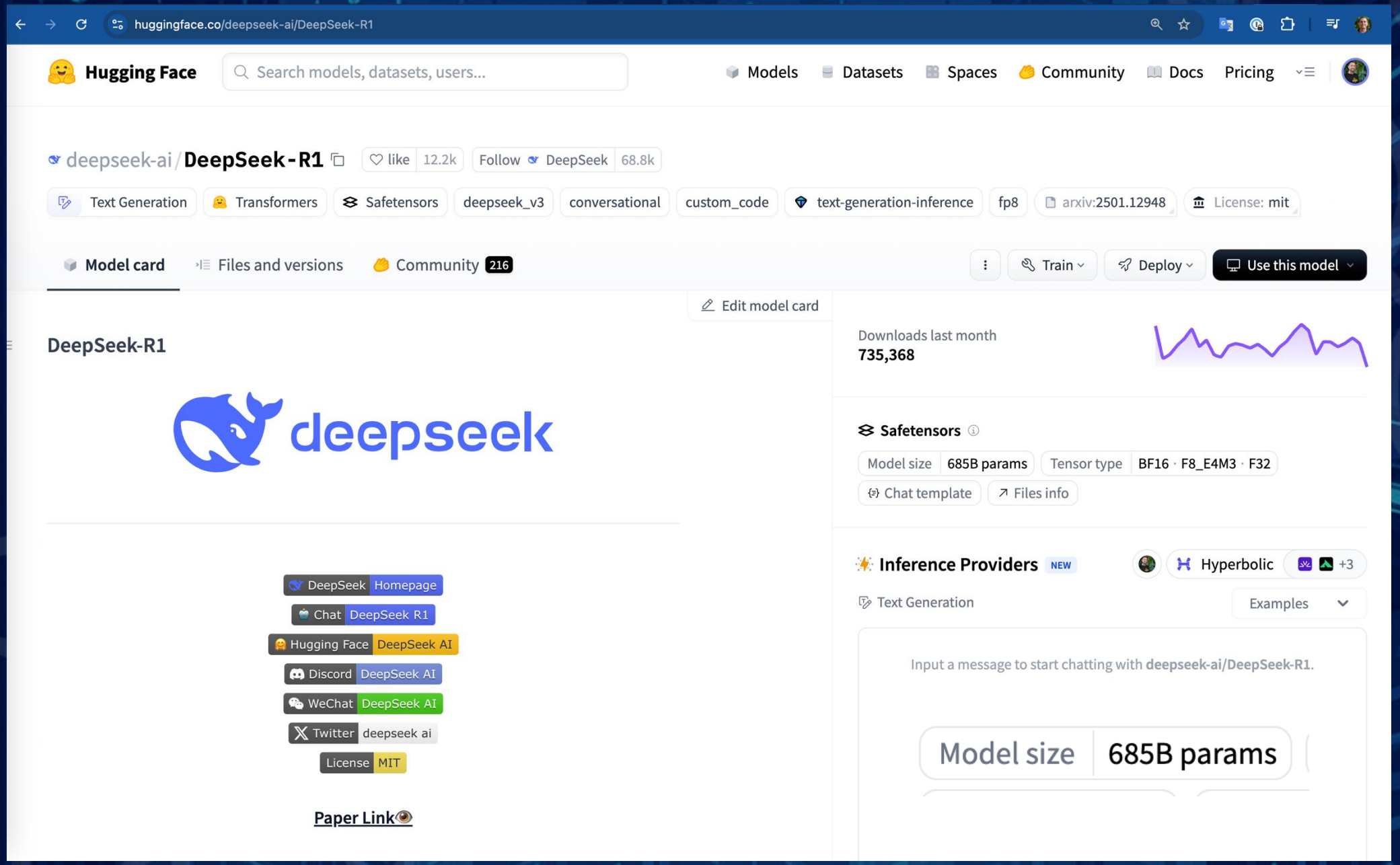
FUTURE TRENDS



- OpenSource models
- SLMs Small Language
- Agentic AI
- Specific hardware
- AI Platforms
- Quantum computing

android
open source project

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
Train
Deploy
Use this model
Edit model card
DeepSeek-R1
deepseek
DeepSeek Homepage
Chat DeepSeek R1
Hugging Face DeepSeek AI
Discord DeepSeek AI
WeChat DeepSeek AI
Twitter deepseek ai
License MIT
Paper Link
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-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 inteligênci...

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 Fe...

Cardápio semanal saudável para ...

Get App NEW

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

Message DeepSeek

DeepThink (R1) Search

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).

fanáticos por tecnologia sac@waz.com.br (31) 2126-6666 (31) 2126-6699

WAZ

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

COMPRAR

POR: R$ 249.999,99
no pix já com 15% de desconto
ou R$ 294.117,64

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.

SLM

Quantization

Distilation

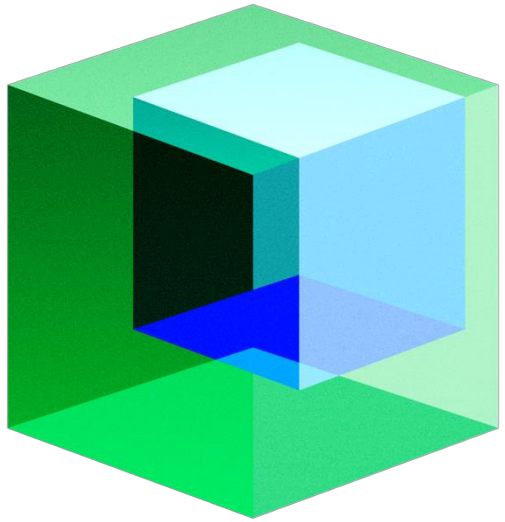
<20b models

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/

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-Guaridan-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

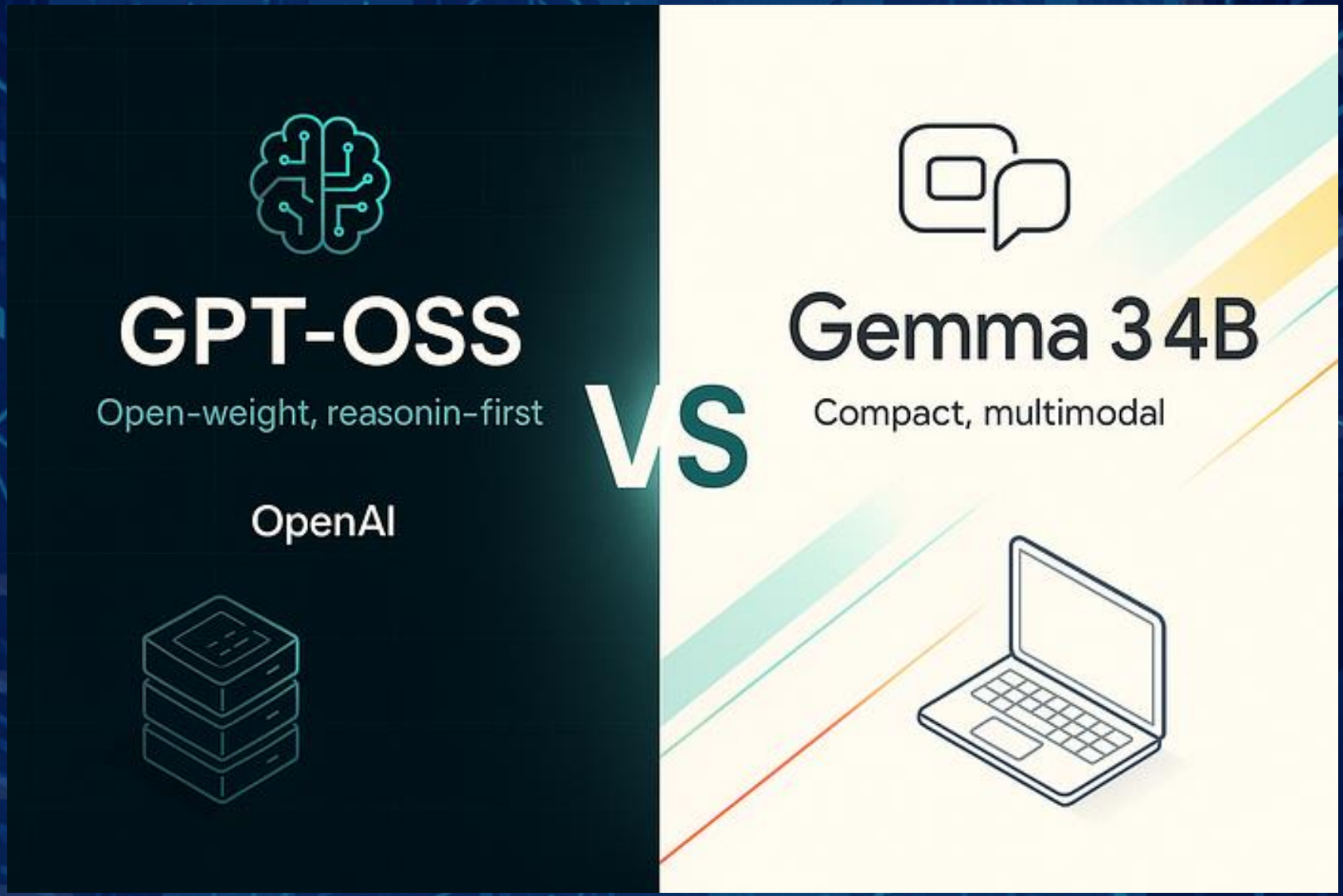
Geospatial models

- Earth

– Granite-EarthObservation-HLS-Biomass
– Granite-EarthObservation-HLS-CanopyHeight
– Granite-EarthObservation-HLS-Landslide

- Weather and climate

– Granite-WeatherClimate-Precip-Downscaling
– Granite-WeatherClimate-WindForecasting



https://www.towardsdeeplearning.com/gpt-oss-vs-gemma-3-two-small-giants-one-big-surprise-ae753d8911e9

OpenAI and Ollama partner to launch gpt-oss

Chat & build with open models

Download

Available for macOS, Windows, and Linux

Why opensource AI is good?

Thought for 7.4 seconds

Open-source AI is good because it unlocks a chain of benefits that individual companies, researchers, students, and even entire nations can share.
Below are the key reasons—grouped into themes—that show why most of the community (and many governments) champion an open-source approach to artificial-intelligence systems.

tability

it matters

Turbo

gpt-oss:20b

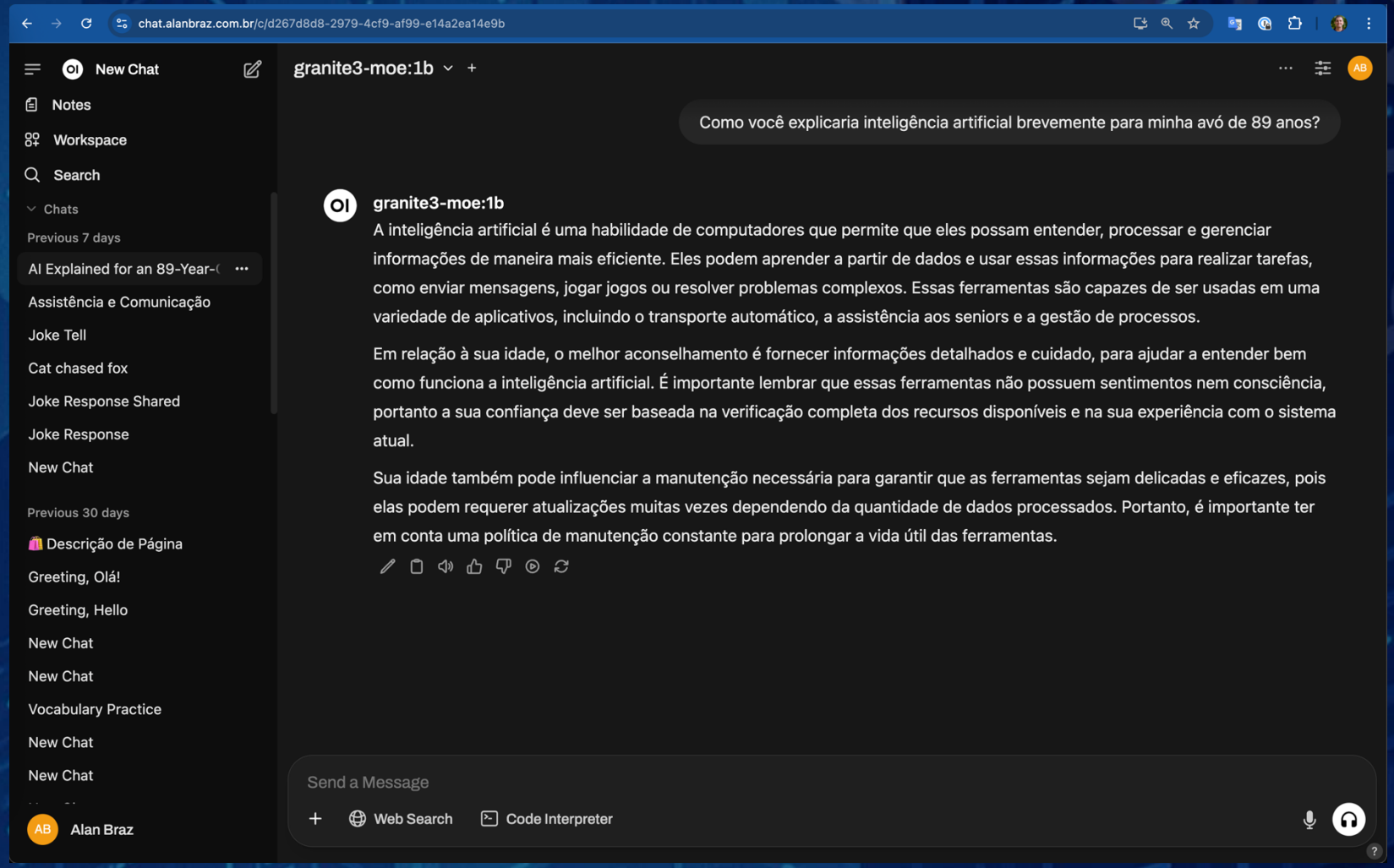
Prc21.txt

resuma a transcrição desse episódio de podcast

gemma3:4b

https://ollama.com/

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



https://chat.alanbraz.com.br

https://www.youtube.com/watch?v=EDb37y_MhRw


Agentes de ia / ai agents / agentic ai

A evolução da IA conversacional

Chatbots

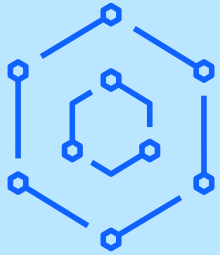
- Baseado em regras
- Determinístico
- Pré-programado

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

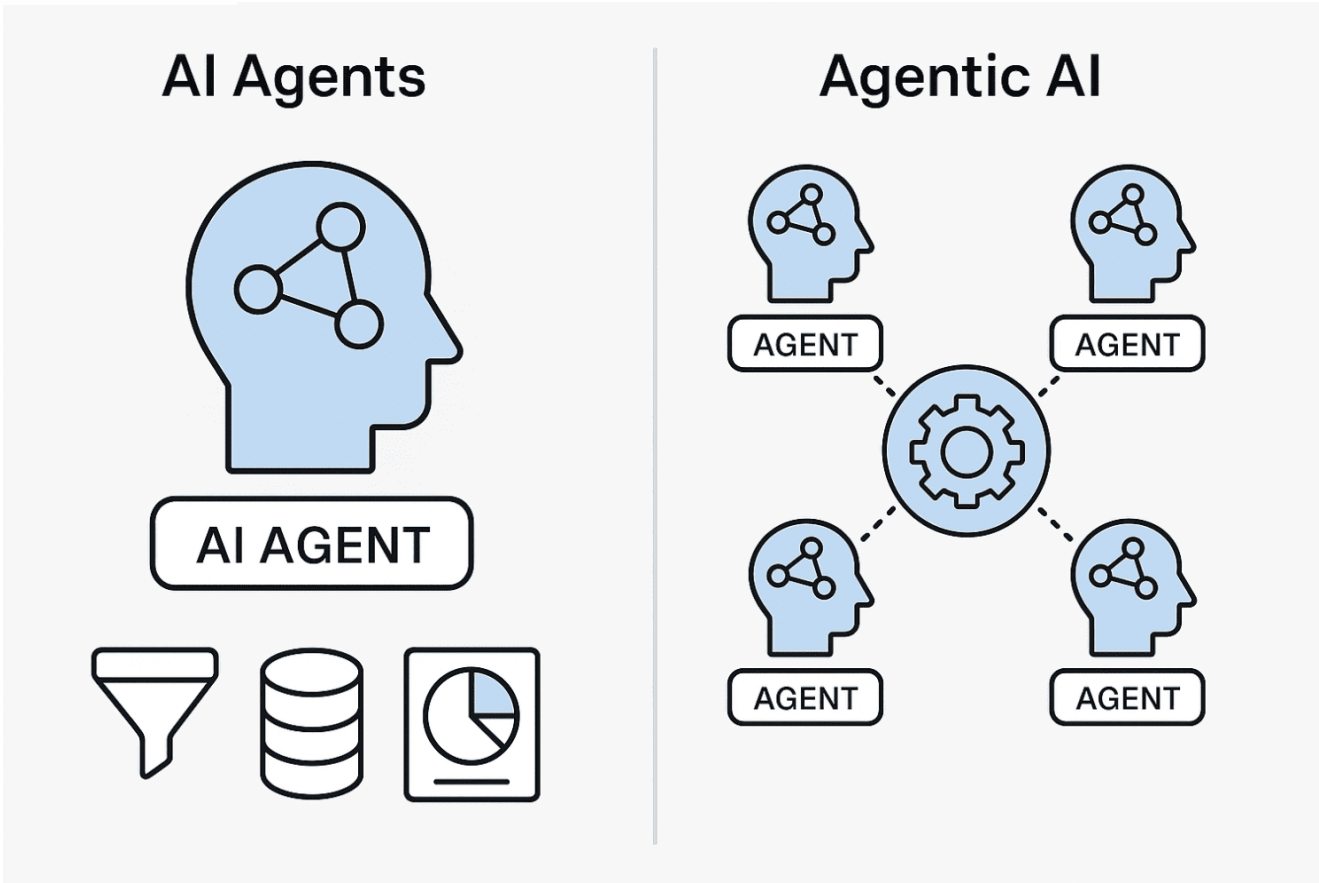
Streamline repetitive tasks

Agentic AI

Manage complex projects

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

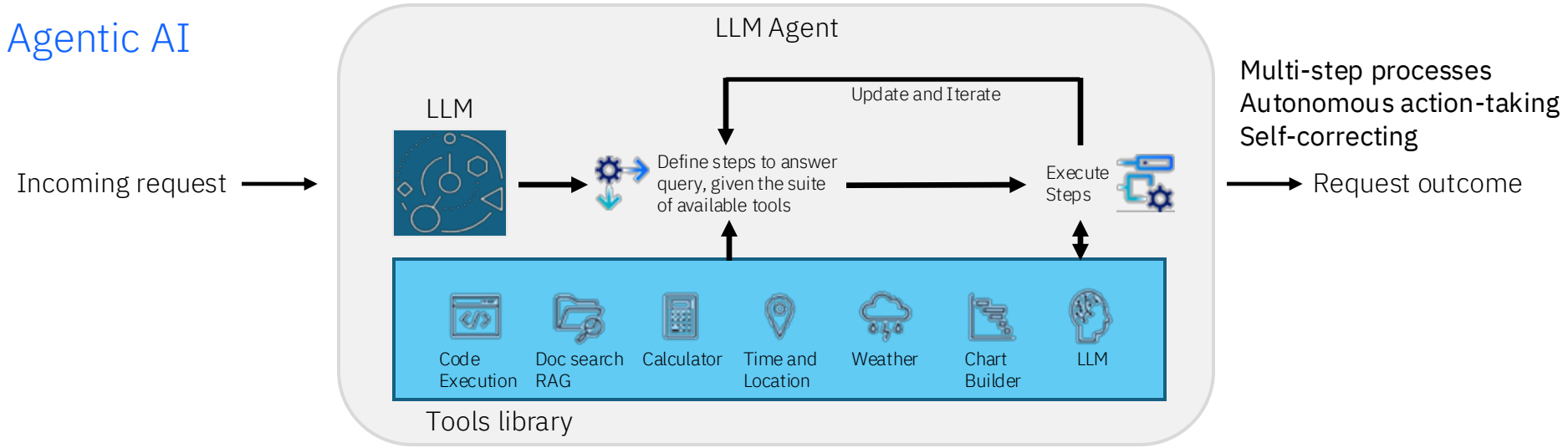
https://theaiinsider.tech/2025/05/19/whats-the-difference-between-ai-agents-and-agentic-ai-new-study-separates-signal-from-noise-in-the-ai-agent-boom/



From Q&A to Agentic AI

Increasing use case complexities

Agentic AI

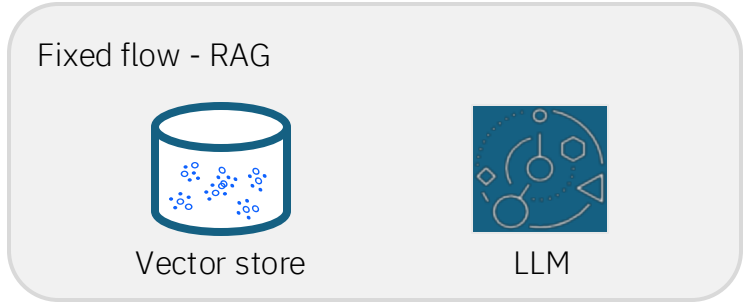


Incoming request

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

Zero/few-shot prompting

Incoming query

Problem-solving
Logical thinking
Pattern matching

Output completion

Capacidades da IA estão crescendo rapidamente

IA que cria

Generative AI

IA que prediz

Machine learning

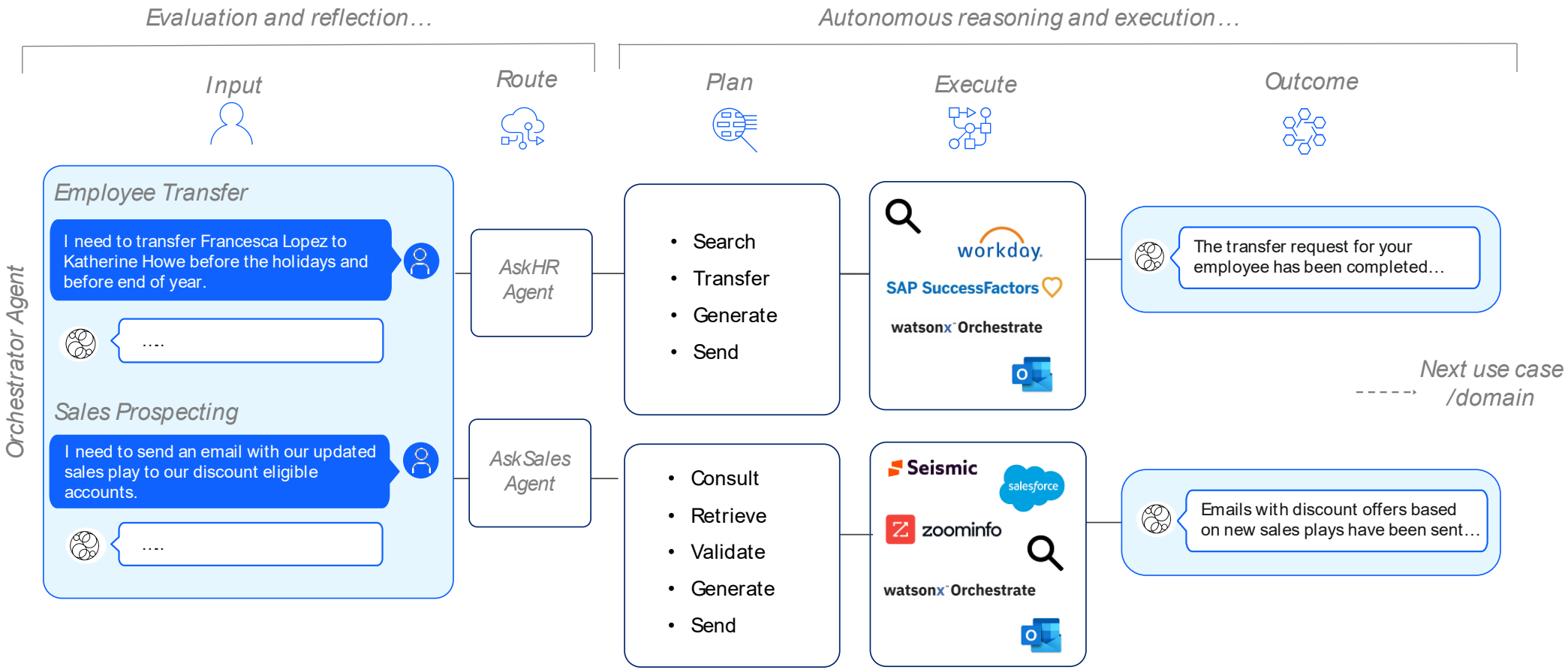
IA que conversa

AI assistants

IA que trabalha

AI agents

Orchestrator Agent



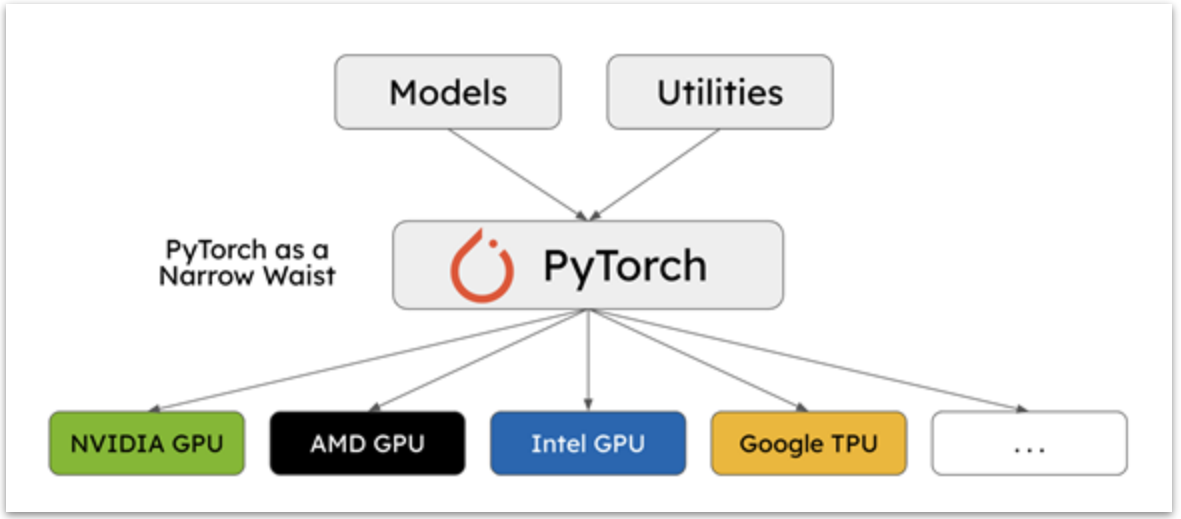
https://www.ibm.com/br-pt/products/watsonx-orchestrate

Hardware Support

NVIDIA GPU

AMD GPU

AWS Neuron



Intel CPU/GPU

Google TPU

Intel Gaudi

IBM Spyre


Microsoft Azure

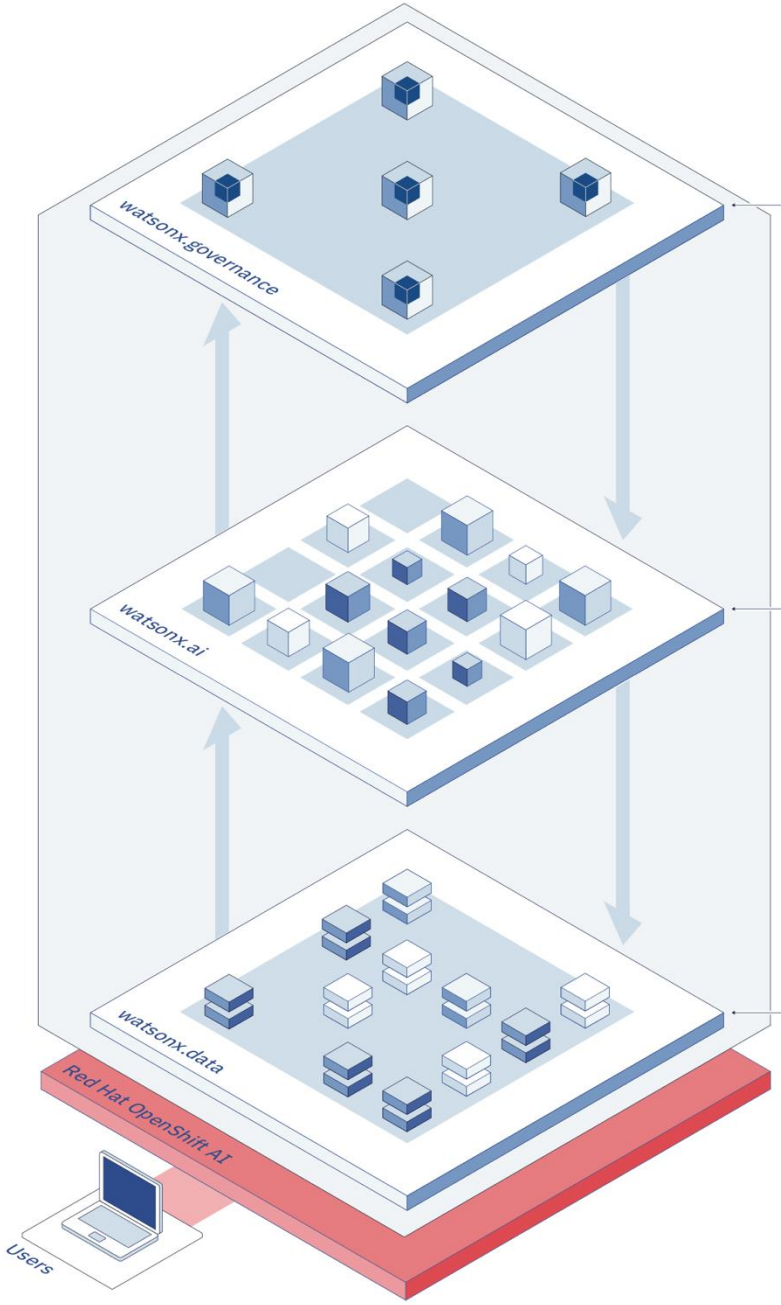

Google AI


aws


OpenAI

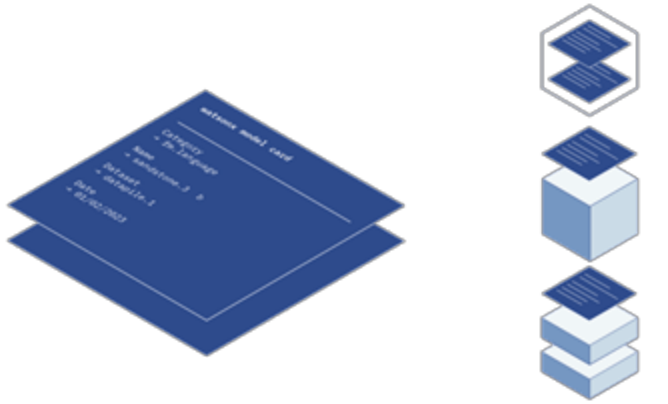

NVIDIA

watsonx, the platform for AI value creators



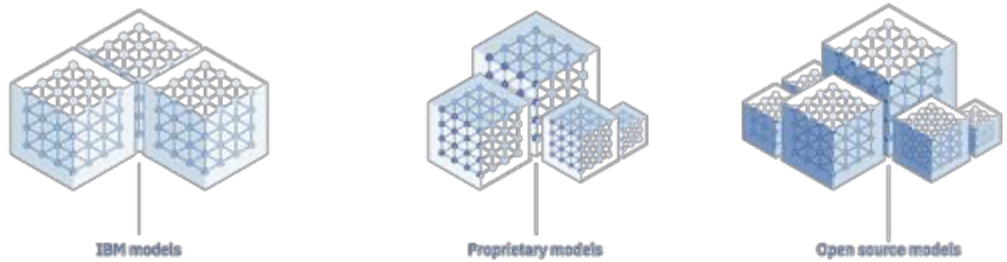
watsonx.governance

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



watsonx.ai

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



watsonx.data

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

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

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.

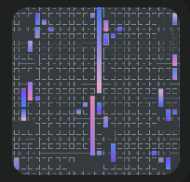
Learn more

https://www.ibm.com/roadmaps/quantum.pdf

qual a previsão da utilidade quântica da IBM?

Perplexity | Images | Sources · 10 | Steps

ibm
The dawn of quantum advantage | IBM...



The Next Platform
IBM Outlines Steps To Verify Claims Of Quantum Advantage

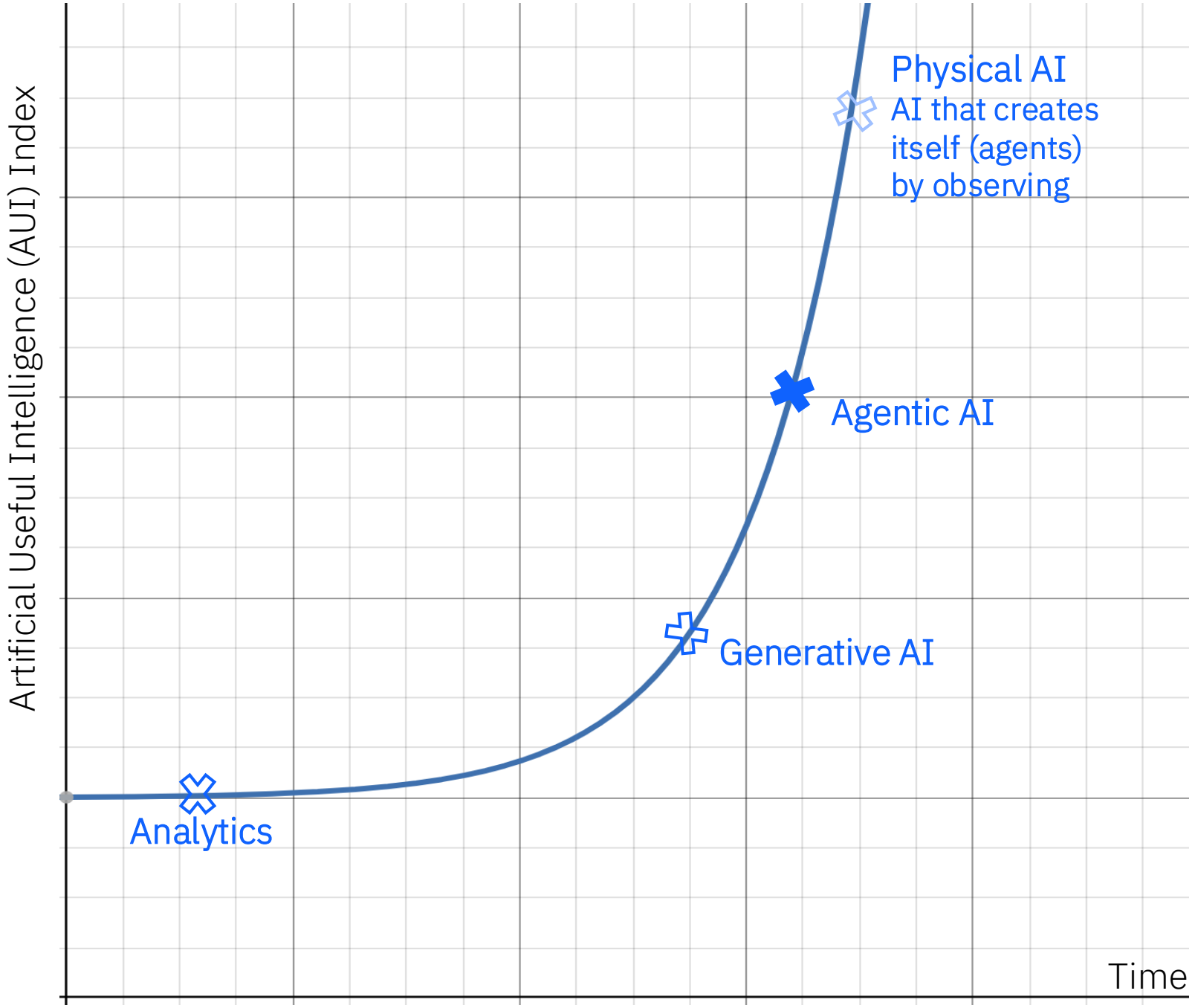
forbes
IBM Promises Enterprise-Ready...

+7 sources

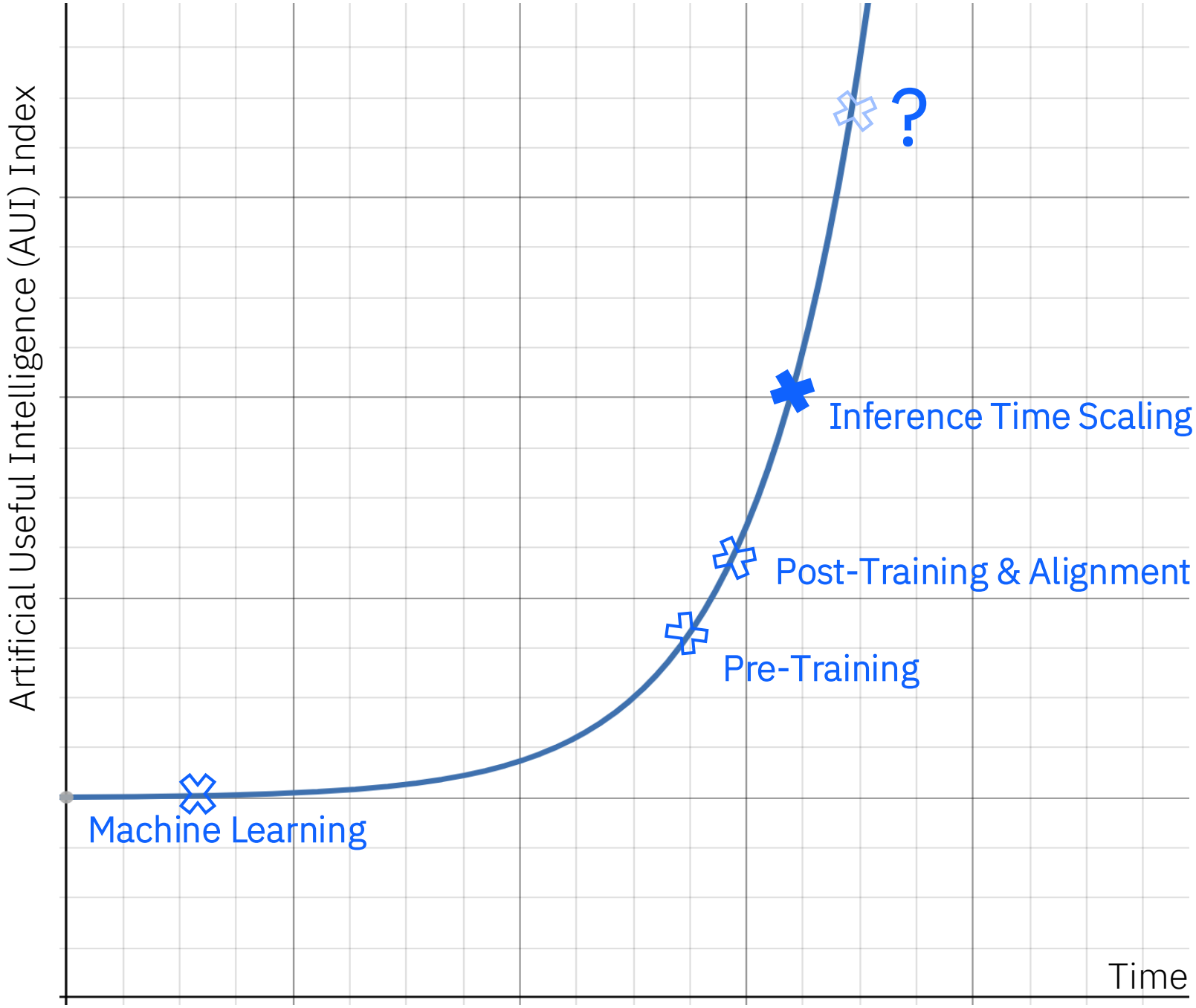
A IBM prevê que a comunidade quântica alcançará a "utilidade quântica" (quantum utility) e começará a demonstrar a "vantagem quântica" (quantum advantage) entre agora e o final de 2026. A utilidade quântica foi um marco alcançado em 2023, onde os computadores quânticos da IBM foram capazes de realizar cálculos confiáveis em uma escala que ultrapassou as simulações clássicas por força bruta de circuitos quânticos. A vantagem quântica, que é superar todos os métodos clássicos, deve surgir com melhorias em hardware, infraestrutura e técnicas de mitigação de erros, e a IBM acredita que será confirmada por um consenso da comunidade científica nos próximos anos.

https://www.perplexity.ai/search/qual-a-previsao-da-utilidade-q-3a4idCrtR4.5cZopYYeSKA

Artificial Intelligence Futures



AI Technology Futures



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

mentoria@pullrecast.dev

Curso de IA OpenSource.

Entre no grupo do WhatsApp