

Desbloqueando a produtividade com watsonx.ai: Como a IA pode revolucionar seu negócio?

Alan Braz
TDC Future 8/12/2023

alanbraz.com.br/tdc-future-2023.pdf



IA Generativa já é realidade?!



da por Inteligência Artificial

Write a poem about Xbox

Xbox, oh Xbox, you're my favorite to
A source of endless fun, a faithful joy
From Halo to Gears, you have it all
A gaming paradise, standing tall

With your sleek design, and wireless
controller
You're the envy of all, a true gamer's
soldier

Your online multiplayer, a world of its
Xbox, you're the king, never overthrown

GITHUB COPILOT: CHAT

monalisa
Write unit tests for this function

GitHub Copilot

```
import datetime
def parse_expenses(expenses_string):
    """Parse the list of expenses and return the list of triples (date, amount, currency)
    Ignore lines starting with #.
    Parse the date using datetime.
    Example expenses_string:
    2023-01-02 -34.01 USD
    2023-01-03 2.59 DKK
    2023-01-03 -2.72 EUR
    """
    expenses = []
    for line in expenses_string.splitlines():
        if line.startswith("#"):
            continue
        date, value, currency = line.split(" ")
        expenses.append(datetime.datetime.strptime(date, "%Y-%m-%d"),
                        float(value),
                        currency)
    return expenses

expenses_data = '''2023-01-02 -34.01 USD
2023-01-03 2.59 DKK
2023-01-03 -2.72 EUR'''
```

AKITANDO AKITAONRAILS AKITAONRAILS AKITAONRAILS

<https://youtube.com/watch?v=CZSGIeRmYn8>

Alan Braz



Research Developer/Software Engineer

- 42 anos, casado, pai de gêmeos (~8)
- Campinas/SP, sofredor do Guarani FC
- Faixa-preta de TaeKwonDo
- Basquete, corrida, vídeo-game
- Bacharel e **Mestre em CC – UNICAMP**
- Já teve a própria empresa de Dev Web 02-05
- **IBM (18)**: Estag, Cons(8), Pesquisa(10)
- Líder do **TLC-BR** 23-24 +3200 membros
- Exp. internacional em Yorktown/NY 17-19
- Criador do CaipiraÁgil (2012) 10x
- Videocast canal.pullrecast.dev

alanbraz.com.br

Career

Client

Code

Content

TECHNICAL LEADERSHIP COMMUNITY

BRAZIL

An IBM Open Innovation Community Chapter

+ 3200
IBMers

145
members

Diversity
Inclusion

AI Automation
Quantum
Security Hybrid Cloud



**PULL
RECAST**

NOVOS EPISÓDIOS SEGUNDAS ÀS 11H33



com Alan Braz
e Matheus Bitencourt

Pull reCast - Desenvolvendo sua Eminência Técnica

@pullreCast 955 subscribers 76 videos

Bem-vindo ao nosso canal! Aqui, Alan Braz e Matheus Bitencourt, dois lide...

pullreCast.dev and 4 more links

INSCREVA-SE

canal.pullreCast.dev

HOME

VIDEOS

SHORTS

LIVE

PLAYLISTS

COMMUNITY

CHANNELS

ABOUT



Episódios completos convidados ► Play all

Bem-vindo ao nosso podcast! Aqui, Alan Braz e Matheus Bitencourt, dois líderes técnicos apaixonados por tecnologia, colaboração e inovação discutem as tendências mais quentes do...



KARINA TRONKOS
@nинаталк, UX Designer,...

Pull reCast - Desenvolvendo sua ...
174 views • 3 days ago



KARINA TRONKOS
De babá à gestora da maior comuni...

Pull reCast - Desenvolvendo sua ...
24 views • 7 days ago



Filipe Dornelas, de estagiário em IA à criador de aplicativo...

Pull reCast - Desenvolvendo sua ...
24 views • 7 days ago



Danne Aguiar,

Confiabilidade além do monitoramento na...

Pull reCast - Desenvolvendo sua ...
10 views • 2 weeks ago



Danne Aguiar, da CM School, e o futuro do...

Pull reCast - Desenvolvendo sua ...
114 views • 4 weeks ago



Emiliano Agazzoni, da CM School, e o futuro do...

Pull reCast - Desenvolvendo sua ...
17 views • 1 month ago

Compactos ► Play all



Compacto Pachi Parra sobre DevRel no GitHub Brasil - P...

Pull reCast - Desenvolvendo sua ...
13 views • 10 days ago



Compacto Filipe Dornelas sobre de estagiário em IA #...

Pull reCast - Desenvolvendo sua ...
21 views • 3 weeks ago



Compacto Pachi Parra sobre de estagiário em IA #...

Pull reCast - Desenvolvendo sua ...
115 views • 1 month ago



Compacto Filipe Dornelas sobre SRE - Pull reCast #17

Pull reCast - Desenvolvendo sua ...
21 views • 1 month ago



Compacto Danne Aguiar, da CM School, e o...

Pull reCast - Desenvolvendo sua ...
21 views • 1 month ago



Compacto Danne Aguiar, da CM School, e o...

Pull reCast - Desenvolvendo sua ...
75 views • 1 month ago

1. Expertise
2. Aprendizado Contínuo
3. Liderança de Pensamento
4. Contribuição
5. Reputação





Inda existe carreira relevante na era da IA generativa?

**Mas afinal,
o que é IA generativa?**

O boom de popularidade começou com
OpenAI e ChatGPT!

NOVOS EPISÓDIOS SEGUNDAS ÀS 11:33 INSCREVA-SE NO CANAL PULLRECAST.DEV

12:48 / 1:13:46

Ainda existe carreira relevante na era da IA generativa? – Pull reCast

Pull reCast - Desenvolvendo sua Eminência Técnica
1.07K subscribers

Subscribed

19 Share Download Clip Save

Impactos do ChatGPT e IA GENERATIVA NOS NEGÓCIOS E NA VIDA

Impactos do ChatGPT e IA GENERATIVA NO RECRUTAMENTO E SELEÇÃO

Impactos do ChatGPT e IA GENERATIVA NA ARQUITETURA DE SOLUÇÕES

Impactos do ChatGPT e IA GENERATIVA EM PROJETOS OPEN-SOURCE

Impactos do ChatGPT e IA GENERATIVA NO DESIGN DE EXPERIÊNCIAS

Impactos do ChatGPT e IA GENERATIVA NA DOCUMENTAÇÃO TÉCNICA

Impactos do ChatGPT e IA GENERATIVA NA SEGURANÇA DA INFORMAÇÃO

Ainda existe carreira relevante na era da IA generativa?

Impactos do ChatGPT e IA GENERATIVA NOS NEGÓCIOS E NA VIDA

Impactos do ChatGPT e IA GENERATIVA NO RECRUTAMENTO E SELEÇÃO

Impactos do ChatGPT e IA GENERATIVA NA ARQUITETURA DE SOLUÇÕES

Impactos do ChatGPT e IA GENERATIVA EM PROJETOS OPEN-SOURCE

Impactos do ChatGPT e IA GENERATIVA NO DESIGN DE EXPERIÊNCIAS

Impactos do ChatGPT e IA GENERATIVA NA DOCUMENTAÇÃO TÉCNICA

Impactos do ChatGPT e IA GENERATIVA NA SEGURANÇA DA INFORMAÇÃO

Ainda existe carreira relevante na era da IA generativa?

- 1 IMPACTOS DO CHATGPT E IA GENERATIVA NOS NEGÓCIOS E NA VIDA 6:50
- 2 IMPACTOS DO CHATGPT E IA GENERATIVA NO RECRUTAMENTO E SELEÇÃO 4:01
- 3 IMPACTOS DO CHATGPT E IA GENERATIVA NA ARQUITETURA DE SOLUÇÕES 4:37
- 4 IMPACTOS DO CHATGPT E IA GENERATIVA EM PROJETOS OPEN-SOURCE 1:42
- 5 IMPACTOS DO CHATGPT E IA GENERATIVA NO DESIGN DE EXPERIÊNCIAS 3:40
- 6 IMPACTOS DO CHATGPT E IA GENERATIVA NA DOCUMENTAÇÃO TÉCNICA 20:24
- 7 IMPACTOS DO CHATGPT E IA GENERATIVA NA DOCUMENTAÇÃO TÉCNICA 8:01
- 8 IMPACTOS DO CHATGPT E IA GENERATIVA NA SEGURANÇA DA INFORMAÇÃO 3:03
- 9 Ainda existe carreira relevante na era da IA generativa? 1:13:47

Inteligência Artificial (IA)

Inteligência humana
simulada por máquinas

IA pode ser definida como uma técnica que permite que as máquinas imitem as funções cognitivas associadas às mentes humanas - as funções cognitivas incluem todos os aspectos da aprendizagem, raciocínio, percepção e resolução de problemas.

1950

Alan Turing publicou o artigo Computing Machinery and Intelligence introduzindo a pergunta: "Máquinas podem pensar?"

1956

John McCarthy cunhou o termo 'inteligência artificial' na primeira conferência de IA no Dartmouth College.

Machine Learning (ML)

Sistemas que aprendem com dados históricos

Aprendizado supervisionado

1Ks a 1Ms de exemplos de dados rotulados

Cada tarefa um modelo

Classificação,
Regressão,
Árvores de decisão

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

Deep Learning (DL)

Técnica de ML que imita a função do cérebro humano

Aprendizado semi- ou não-supervisionado

1Ms de exemplos de dados rotulados e não rotulados

Imagens e sons

"Aprendizado Profundo": DL é um subconjunto de ML, usando várias camadas de redes neurais, que são nós interconectados, que trabalham juntos para processar informações. DL é adequado para aplicativos complexos, como reconhecimento de imagem e fala.

Foundation Model

Sistemas de IA generativos

Aprendizado não supervisionado
Bilhões de dados não rotulados
Grandes modelos de linguagem LLM
Modelo "base" pré-treinado de forma genérica
GPT-3: 1024 GPUs, 34 dias, \$4.6M

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

Fine-tuning:

Zero-shot prompting: No labeled data

Few-shot prompting: 1-10 labeled data

Data-driven tuning: 100s – 1000s labeled

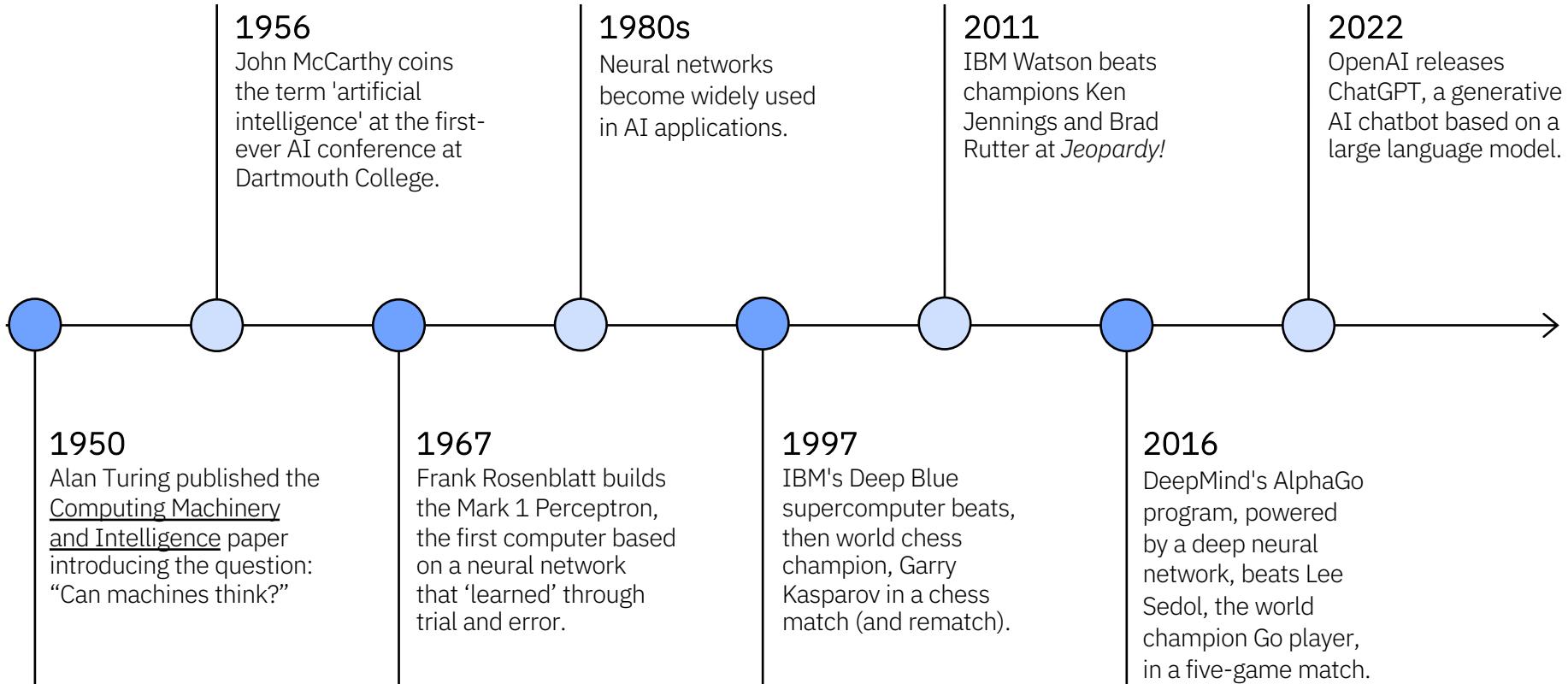


NOVOS EPISÓDIOS SEGUNDAS ÀS 11:33

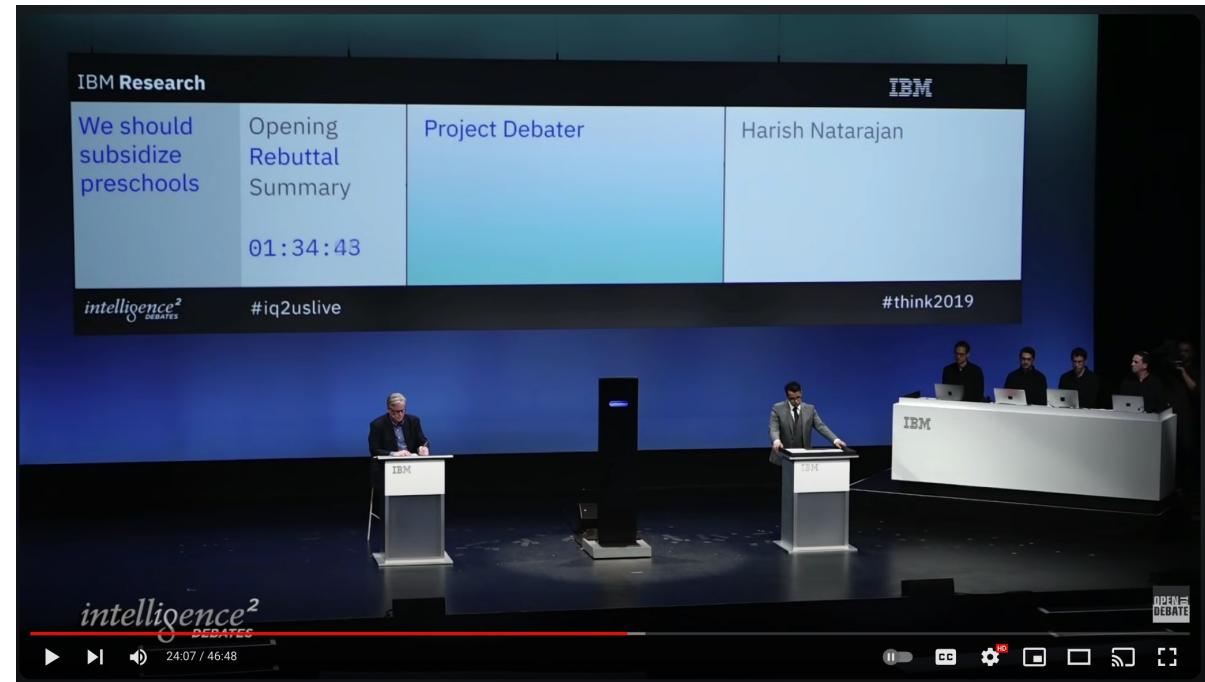


INSCREVA-SE NO CANAL.PULLRECAST.DEV

AI milestones



Project Debater 2019



<https://research.ibm.com/interactive/project-debater/film/>
https://youtube.com/watch?v=3_yy0dnIc58

Generative AI and traditional AI

Both traditional AI and generative AI are useful for enterprises.

Neither replaces the other, generative AI [opens new possibilities](#)

Generative AI

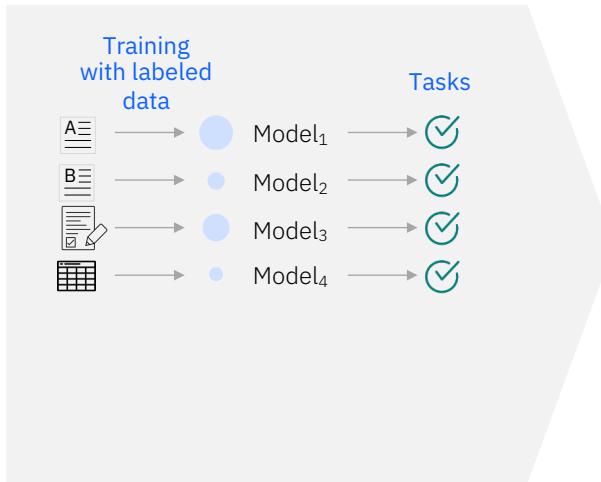
- Foundation models trained with unlabeled data
- Unsupervised
- Trained on very big data sets
- No specific task
- Transferable
- Works well for general tasks and can improve for specific tasks with less training
- Need to monitor bias and drift

Traditional AI

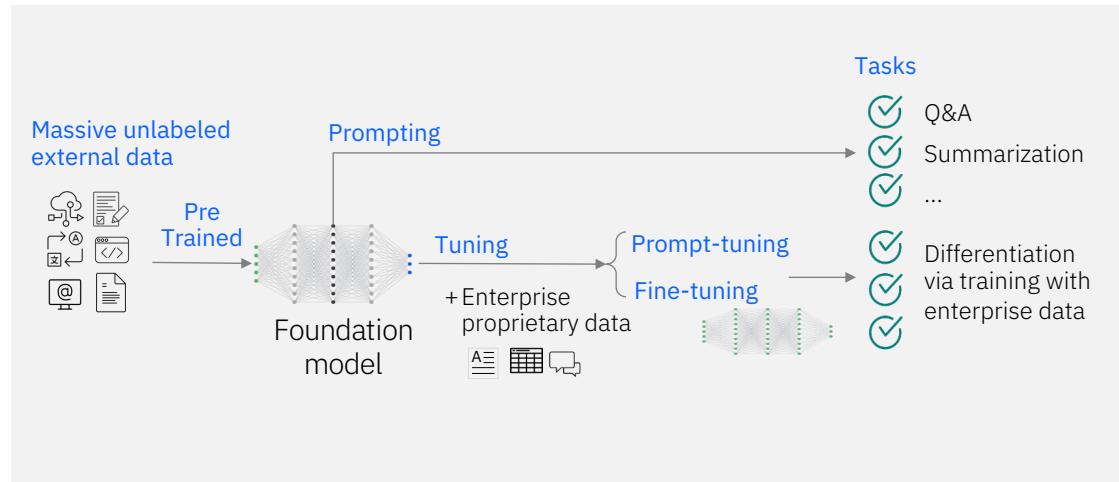
- Traditional Machine learning (ML/AI) model trained with “labeled” data
- Training is supervised
- Trained on proper, large data sets
- Trained for a specific task
- Does not transfer well to other tasks
- A tuned model can be very efficient for the specific task it was designed for
- Need to monitor bias and drift

Foundational models enable a new paradigm of data-efficient AI development – generative AI

Traditional AI models



Foundation Models



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

- Rapid adaptation to multiple tasks with small amounts of task-specific data
- Pre-trained unsupervised learning

Impact of generative AI

The speed, scope, and scale of generative AI impact is unprecedented

Massive early adoption

80%
of enterprises are working with or planning to leverage foundation models and adopt generative AI

Sources: Statista; Reuters; Goldman Sachs; IBM Institute for Business Value; Gartner. Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners

Broad-reaching and deep impact

Generative AI could raise global GDP by 7% within 10 years

Critical focus of AI activity and investment

Generative AI expected to represent 30% of overall market by 2025

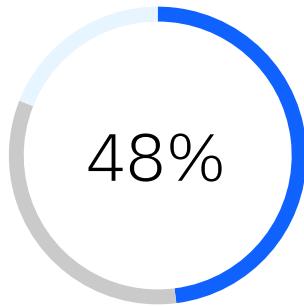
Most common generative AI tasks implemented today

<p>Summarization</p> <p>Transform text with domain-specific content into personalized overviews that capture key points.</p> <p><i>Conversation summaries, insurance coverage, meeting transcripts, contract information</i></p>	<p>Classification</p> <p>Read and classify written input with as few as zero examples.</p> <p><i>Sorting of customer complaints, threat and vulnerability classification, sentiment analysis, customer segmentation</i></p>	<p>Generation</p> <p>Generate text content for a specific purpose.</p> <p><i>Marketing campaigns, job descriptions, blog posts and articles, email drafting support</i></p>
<p>Extraction</p> <p>Analyze and extract essential information from unstructured text.</p> <p><i>Medical diagnosis support, user research findings</i></p>	<p>Question-answering</p> <p>Create a question-answering feature grounded on specific content.</p> <p><i>Build a product specific Q&A resource for customer service agents.</i></p>	

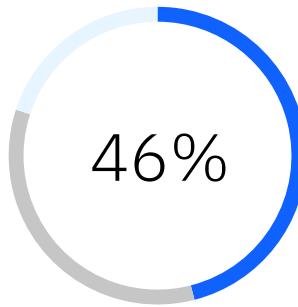
Generative AI adoption considerations, inhibitors and fears

80% of business leaders see at least one of these ethical issues as a major concern

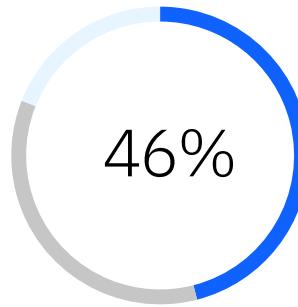
Explainability



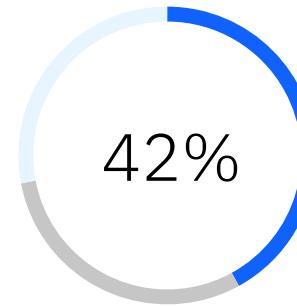
Ethics



Bias



Trust



Believe decisions made by generative AI are not sufficiently **explainable**.

Concerned about the safety and **ethical** aspects of generative AI.

Believe that generative AI will propagate established **biases**.

Believe generative AI cannot be **trusted**.

█ Agree █ Neutral █ Disagree

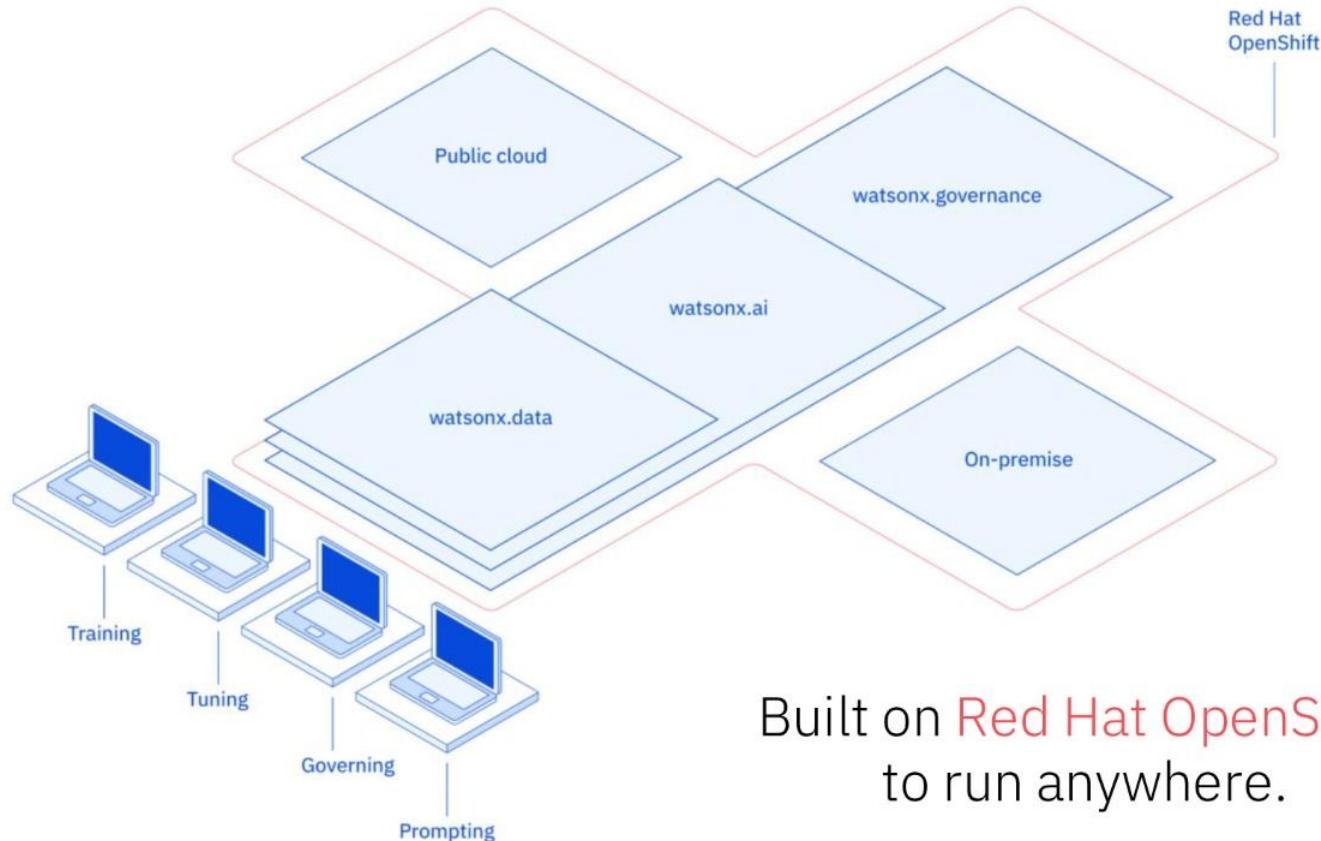
Enterprises need more than an AI solution - they need a comprehensive and sound strategy for generative AI.

Introducing...

watsonx.ai

How to scale AI across
the business

Putting AI to work on Hybrid Cloud



Built on Red Hat OpenShift.
to run anywhere.



What IBM offers

The platform
for AI and data

watsonX

Scale and
accelerate the
impact of AI with
trusted data.

watsonX.ai

Train, validate, tune and
deploy AI models

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

Scale AI workloads, for all
your data, anywhere

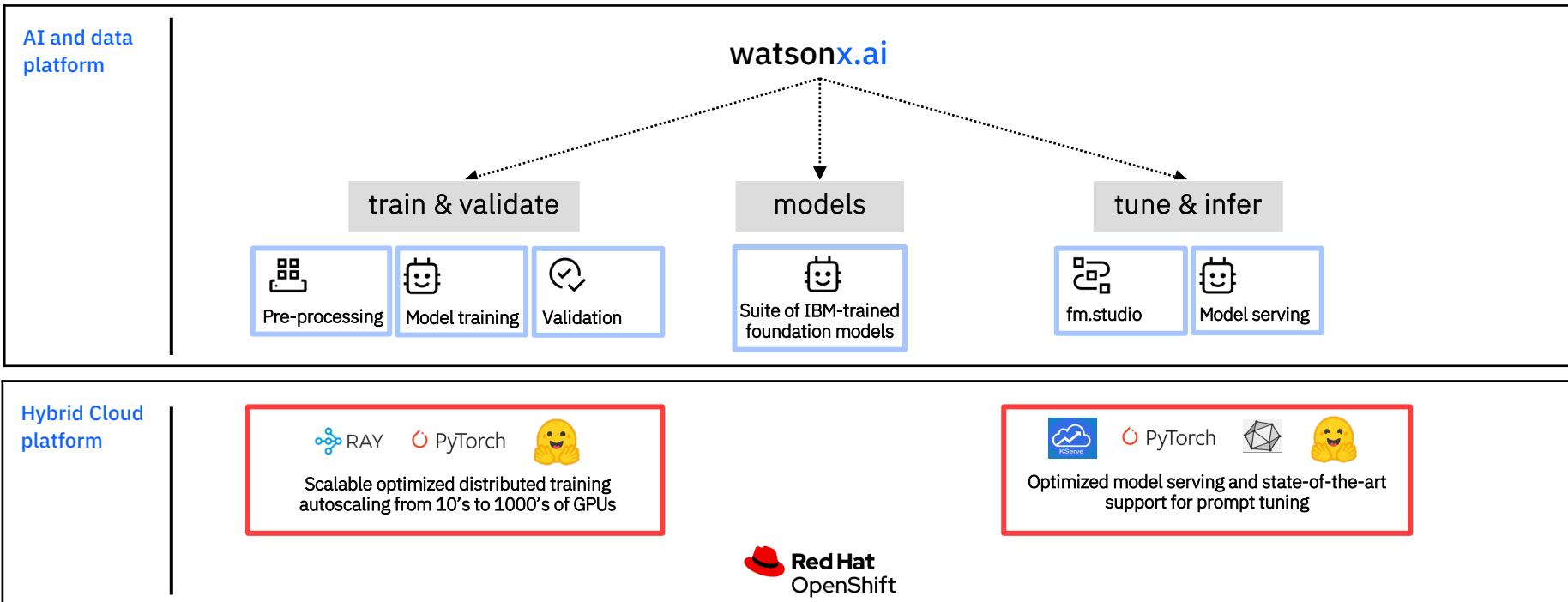
Fit-for-purpose data store, built on
an open lakehouse architecture,
supported by querying, governance
and open data formats to access
and share data.

watsonX.governance

Accelerate responsible,
transparent and explainable AI
workflows

End-to-end toolkit for AI
governance across the entire model
lifecycle to accelerate responsible,
transparent, and explainable AI
workflows

watsonx.ai drives new IBM innovation and Red Hat OpenShift enhancements for AI workloads



What IBM offers

watsonx: Foundation Model Library

Explore the different foundation models offered in **watsonx** to cover a range of enterprise use cases

IBM Foundation Models

- Transparent Pre-Training on IBM's trusted Data Lake
- We've built and carefully vetted one of the largest business relevant data sets to train our models based on principles of trust and transparency
- Full, auditable data lineage available for any IBM Model

Slate (encoder only) Natural Language Processing Models

- multilingual, distilled (153 million)
- Fine tuning, via notebooks and APIs, required to support extract and classify language tasks
- optimized by [Intel OneAPI library](#)

Granite (decoder only) **NEW**

- Natural Language Processing Models ([watsonx.ai](#))
- Code Models ([Watson Code Assistant](#))
- 13b parameter models fit on V100-32GB GPU

Open-source large language models

- Models sourced from Hugging Face 
- Models sourced from partnerships and vetted for inclusion on [watsonx.ai](#)

- **flan-ul2** (20b parameters, encoder/decoder);
Supports Q&A, Generate, Extract, Summarize, Classify tasks

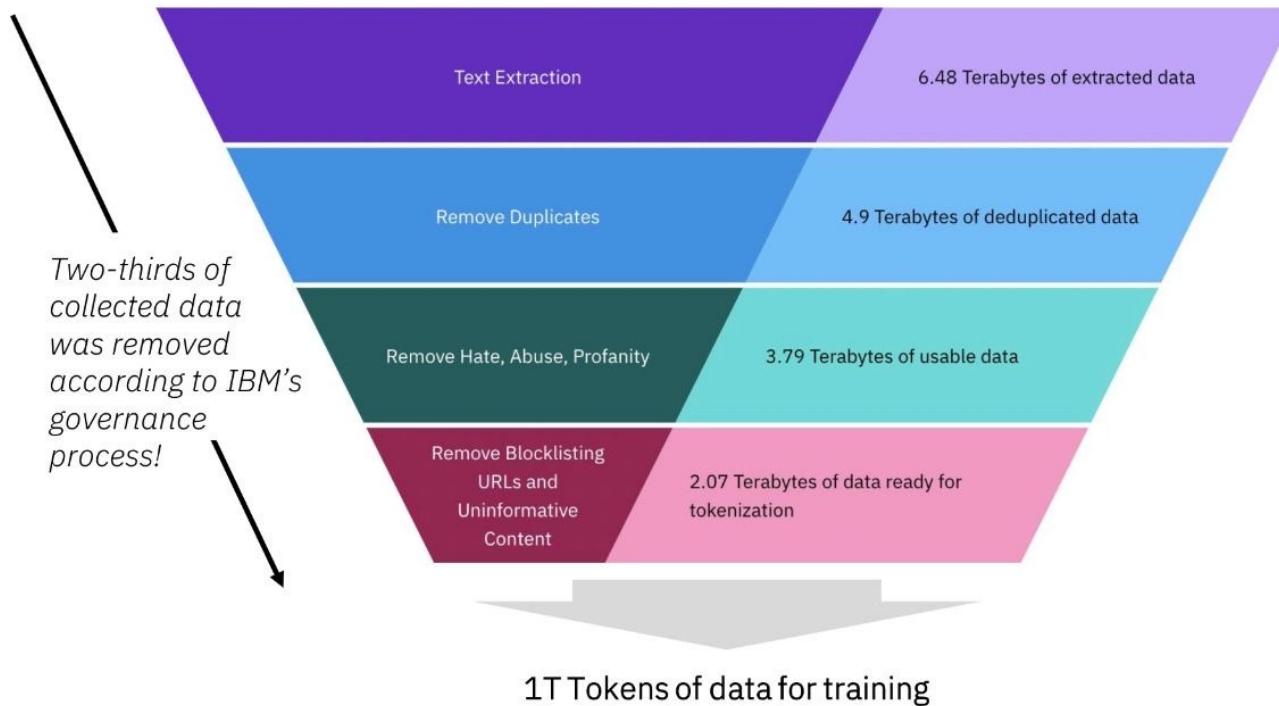
- **gpt-neox** (20b parameters, decoder only);
Supports Q&A and Generate tasks

- **starcoder** (15.5b parameters, decoder only)
Supports code tasks

- **LLama2** (70b parameters, decoder only, chat variant) **NEW**

... And more!

Granite.13b: Training data governance funnel



watsonx.ai: Prompt Lab

Experiment with foundation models and build prompts

Interactive prompt builder

Includes prompt examples for various use cases and tasks

Experiment with different prompts, save and reuse older prompts, use different models and vary different parameters

Experiment with zero-shot, one-shot, or few-shot prompting to get the best results

Experiment with prompt engineering

Choice of foundation models to use based on task requirements

Prevent the model from generating repeating phrases

Number of min and max new tokens in the response

Stop sequences – specifies sequences whose appearances should stop the model

The screenshot shows the IBM WatsonX Prompt Lab interface. At the top, there's a navigation bar with 'IBM watsonx', a search bar, and various account and location options. The main area is titled 'Prompt Lab' and shows a workspace named 'New (unsaved)'. On the left, a sidebar titled 'Sample prompts' lists several options: 'Summarization', 'Earnings call summary' (selected), 'Meeting transcript summary' (highlighted in blue), 'Scenario classification', 'Sentiment classification', 'Marketing email generation', 'Thank you note generation', 'Named entity extraction', 'Fact extraction', and 'Question answering'. The main content area has tabs for 'Structured' and 'Freeform'. Under 'Set up', there's an 'Instruction (optional)' field with the placeholder 'Write a short summary for the meeting transcripts.' Below it is an 'Examples (optional)' section with a table:

Transcript	Summary
00:00 [John] I wanted to share an update on project X today. 00:15 [John] Project X will be completed this week 00:35 [Jane] I heard from customer Y today, and they agree...	John shared an update that project X will be completed this week and will be purchased by customers Y and Z.
00:00 [Jane] The goal today is to agree on a design solution. 00:12 [John] I think we should consider choice 1. 00:40 [Joe] Choice 2 has the advantage that it will take les...	Jane, John, and Joe decided to go with choice 2 for the design solution because it will take less time.

At the bottom, there's a 'Try' section with a 'Test your prompt' input field containing '1' and a summary table:

Transcript	Summary
1	John Doe 00:00:01.415 --> 00:00:20.675

Below the summary table, it says 'Time running: 80 out of 40966.98 second' and a large blue 'Generate' button.

watsonx.ai: Tuning Studio

Tune your foundation models with labeled data

Prompt tuning

Efficient, low-cost way of adapting an AI foundation model to new downstream tasks

Tune the prompts with no changes to the underlying base model or weights

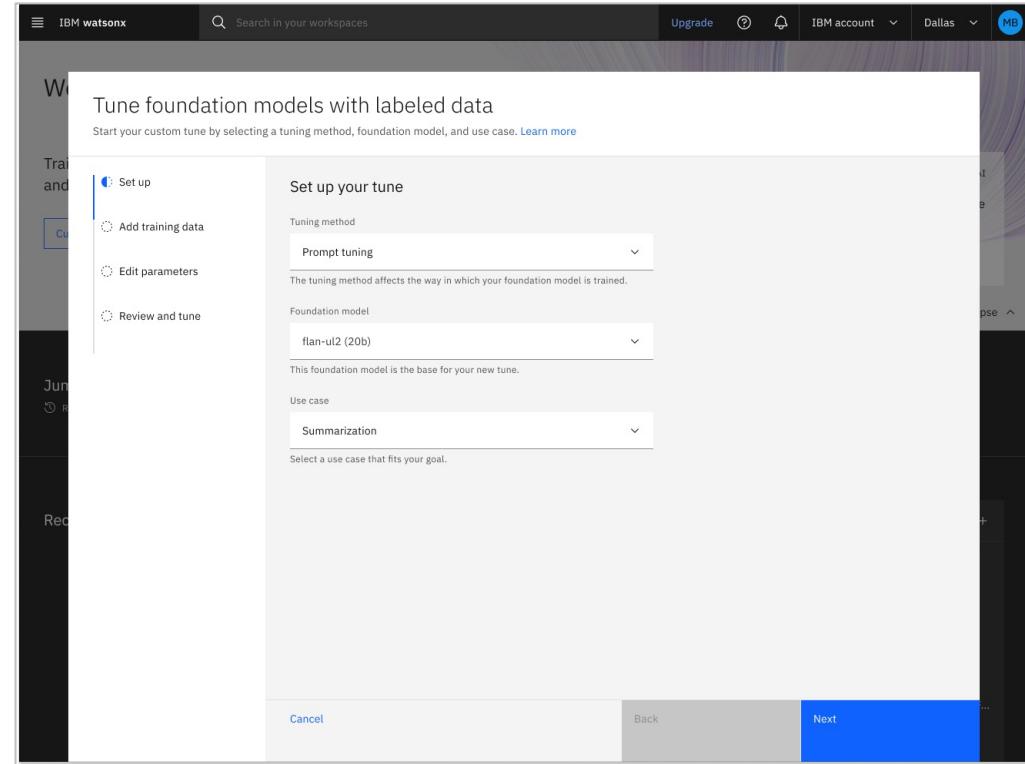
Unlike prompt engineering, prompt tuning allows clients to further train the model with focused, business data

Task support in the Tuning Studio

Models support a range of Language Tasks: Q&A, Generate, Extract, Summarize, Classify

Requires a small set of labelled data to perform specialized tasks

Can achieve close to fine-tuning results without model modification, at a lower cost to run



watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

Build experiments quickly and enhance training by optimizing pipelines and identifying the right combination of data

AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

Use predictions to optimize decisions, create and edit models in Python, in OPL or with natural language

Integrated visual modeling

Prepare data quickly and develop models visually to help visualize and analyze enterprise data to identify patterns and trends, explore opportunities, and make informed, insightful business decisions

- Uncover correlations
- Insight for hypotheses
- Find relationships and connections within the data

The screenshot shows the WatsonX AI Studio interface. At the top, there's a search bar with the placeholder "Search for a task or tool" and a dropdown menu "All coding types". Below the search bar, there are two main sections: "Prepare data" and "Work with models".

Prepare data section:

- with Connections:** Connect to a data source (AI icon), Prepare and visualize data (AI icon).
- with Data Refinery:** (empty)

Work with models section:

- with Prompt Lab:** Experiment with foundation models and build prompts (AI icon).
- with AutoAI:** Build machine learning models automatically (AI icon).
- with SPSS Modeler:** Build models as a visual flow (AI icon).
- with Jupyter notebook editor:** Work with data and models in Python or R notebooks (AI icon).
- AI icon:** Solve optimization problems (AI icon).
- AI icon:** Train models on distributed data (AI icon).
- AI icon:** Write R notebooks and script (AI icon).

What IBM offers

watsonx.ai is helping companies custom build AI solutions to suit their specific needs.



Leveraged watsonx.ai foundation models to train their AI to create tennis commentary. [Generated informative and engaging video clip narrations](#) for fans with varied sentence structures and vocabulary.



SAMSUNG SDS

Exploring watsonx.ai's generative AI capabilities for new solutions such as SDS's Zero Touch Mobility to [deliver unprecedented product innovations](#) to improve client experience.



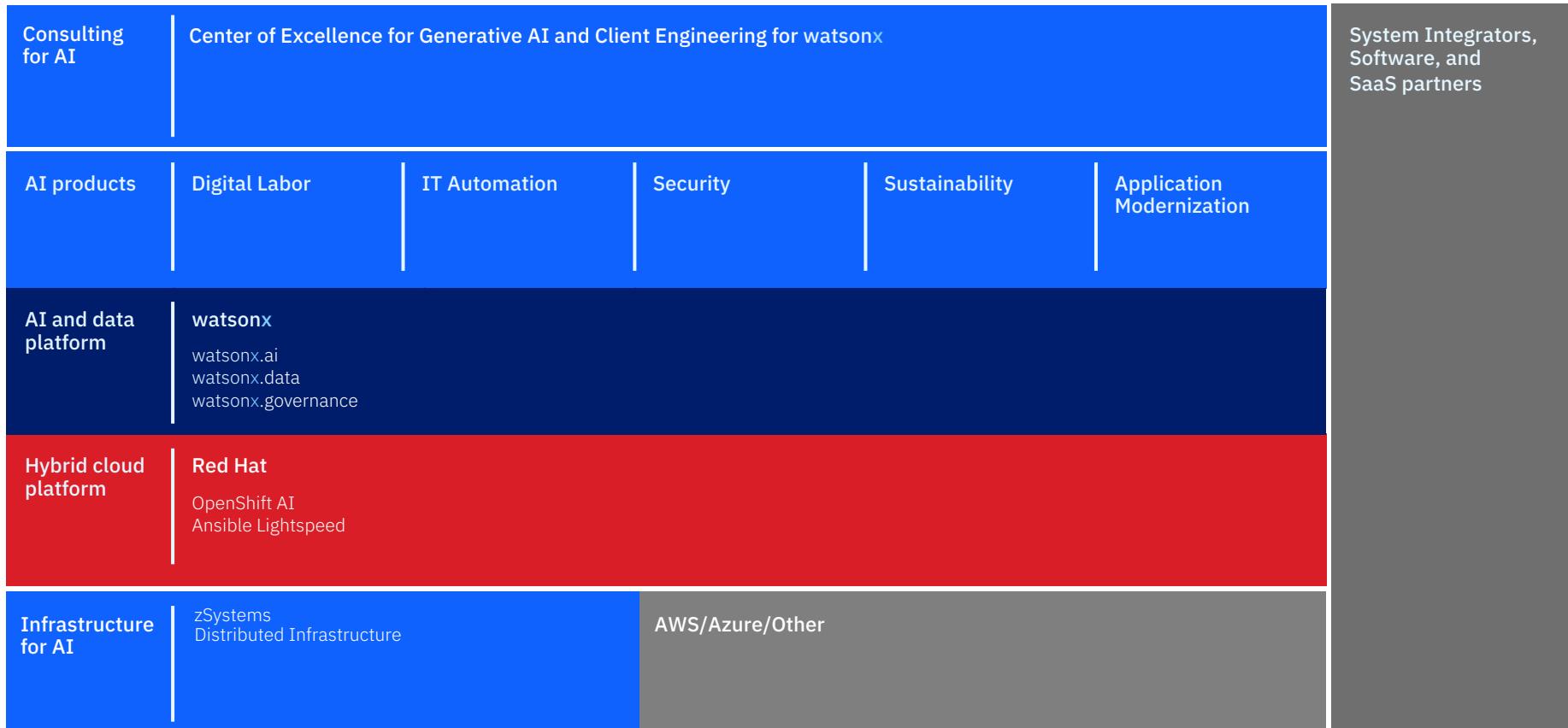
Using watsonx.ai to [slash delivery time from 3-4 months down to 3-4 weeks](#) for many customer care use cases.



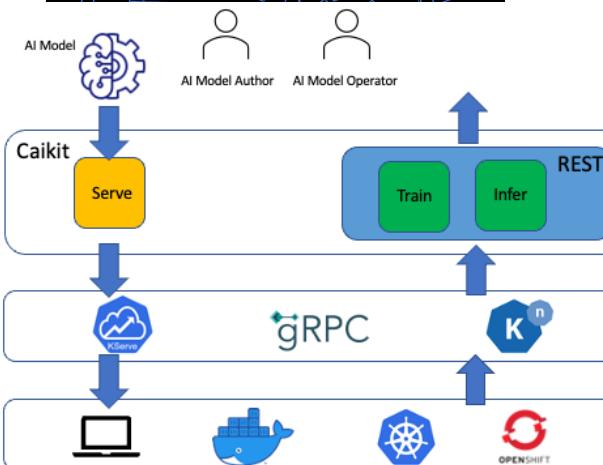
An early adopter of generative AI, has been exploring watsonx.ai to improve [content discoverability, summarization and classification of data](#) to enhance productivity.

What IBM offers

IBM approach for AI: Unleash the intelligence in your business



A cloud-native, open-source stack for accelerating foundation model innovation

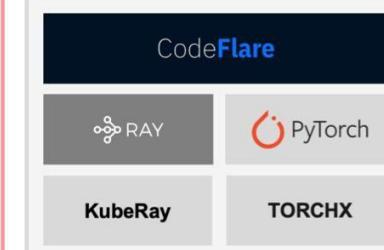


<https://github.com/caikit/caikit>

Training and validation



Workflows



Tuning and inference



Domain specific APIs



Red Hat OpenShift AI



Multi-NIC CNI



IBM Cloud



Why IBM?

Open

IBM's AI is based on the best open technologies available

Trusted

IBM's AI is transparent, responsible, and governed

Targeted

IBM's AI is designed for enterprise and targeted at business domains

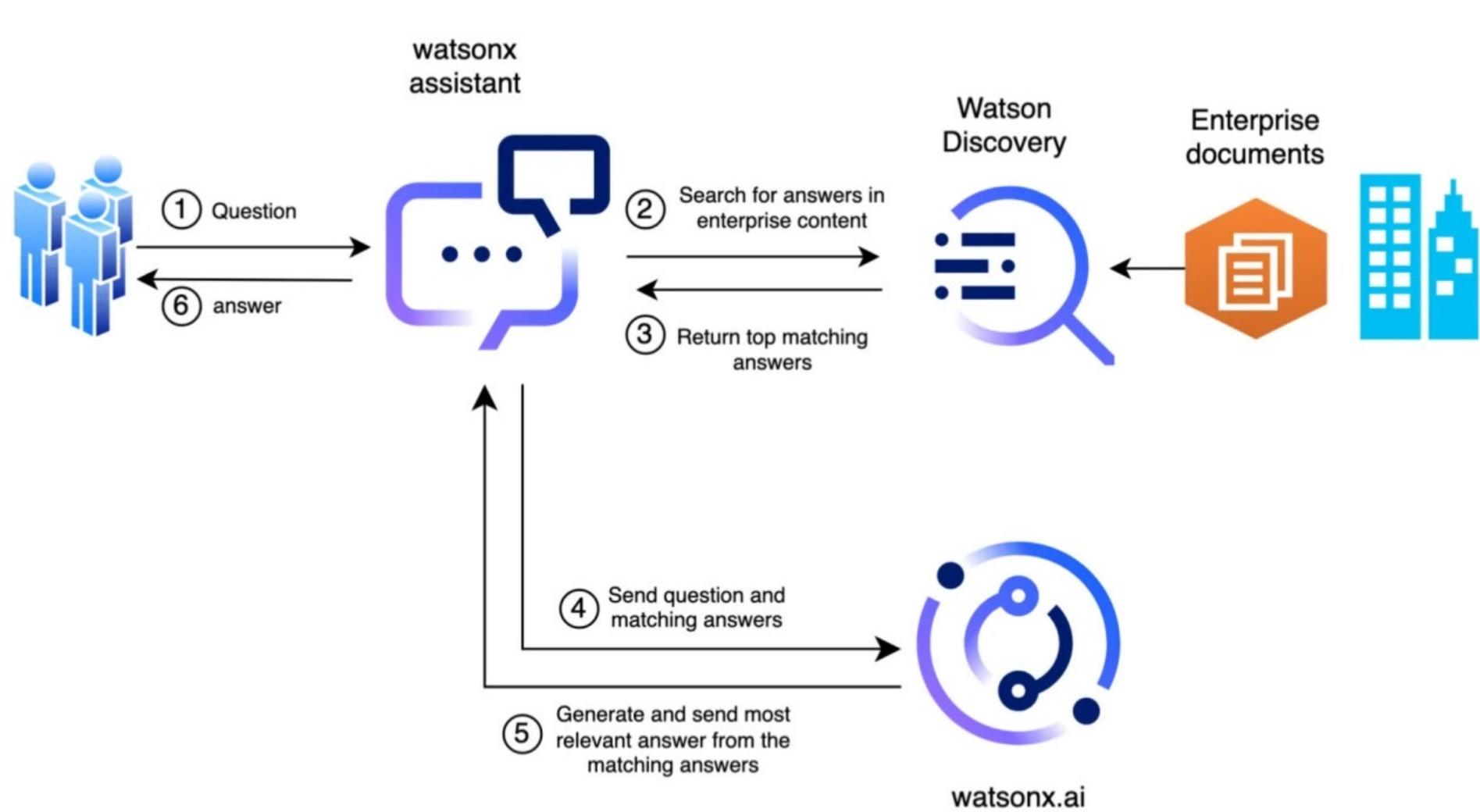
Empowering

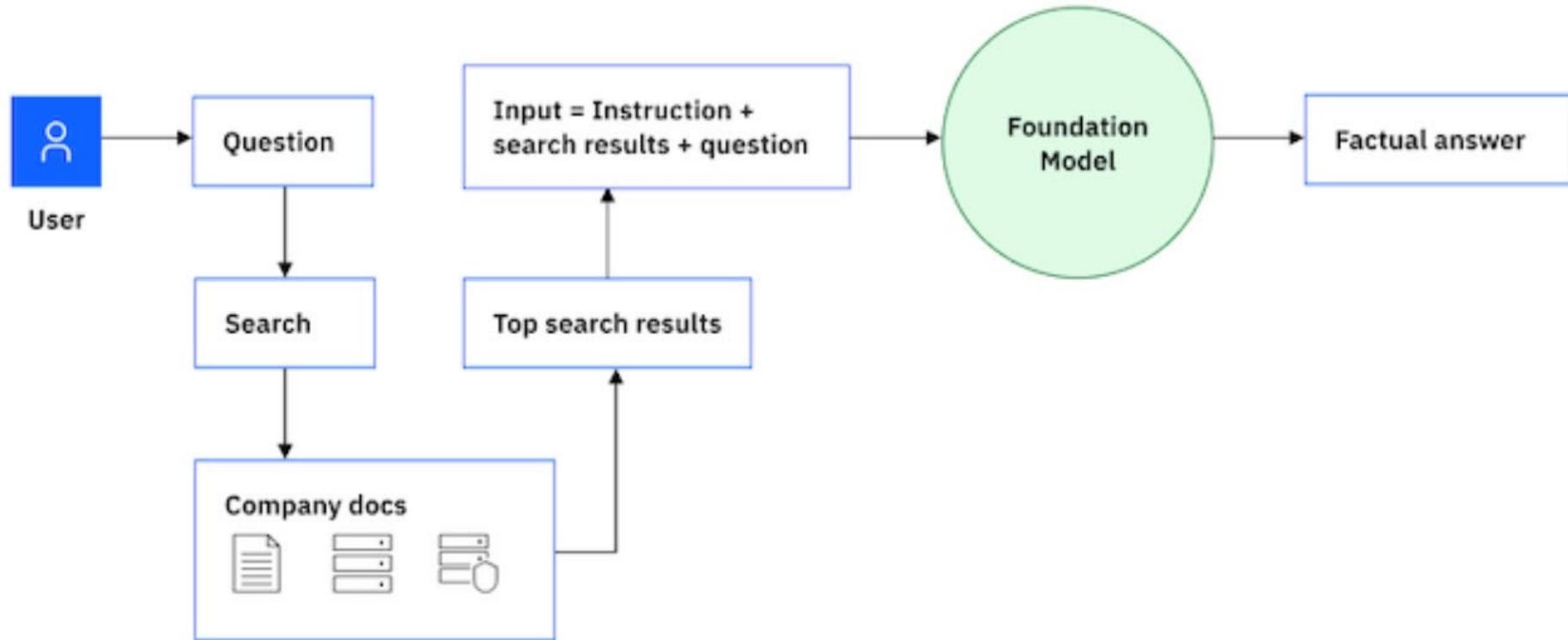
IBM's AI is for value creators, not just users

Retrieval Augmented Generation

RAG

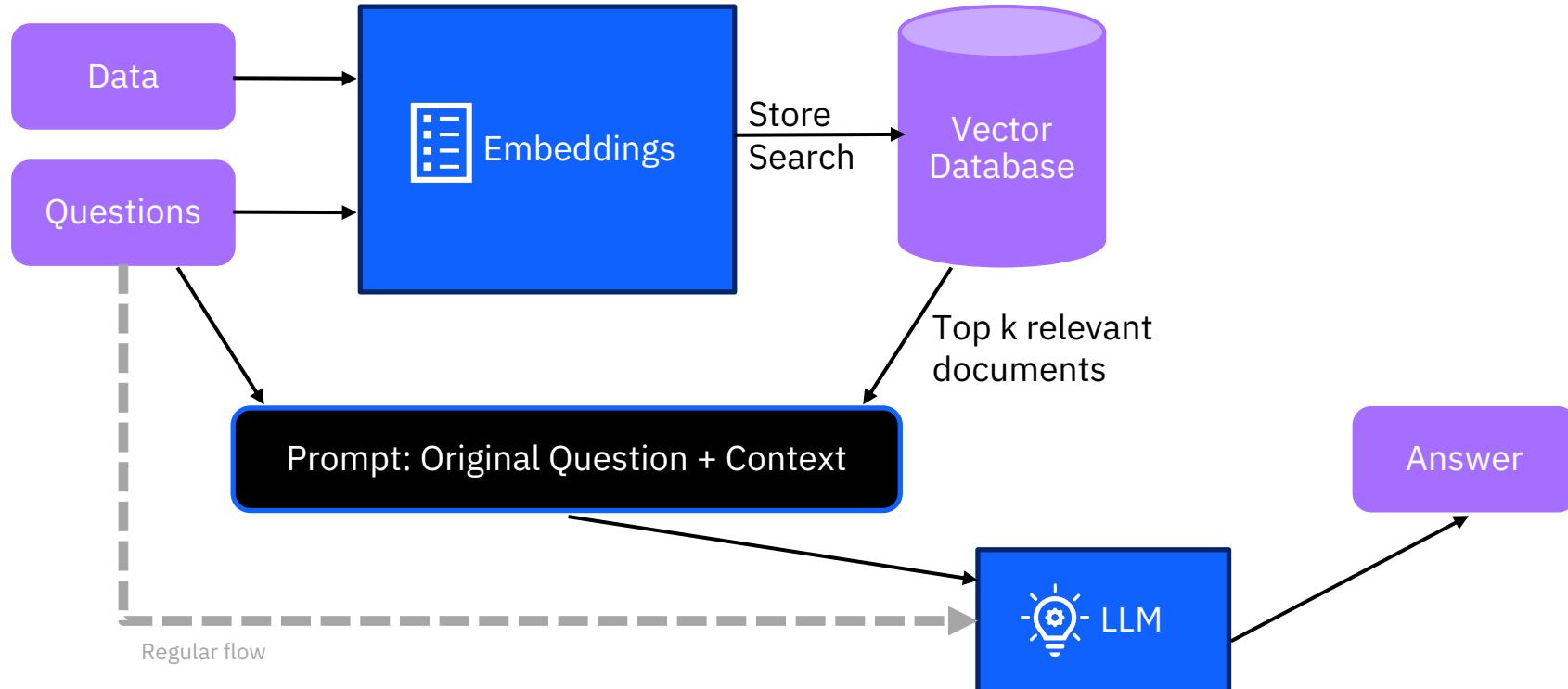






Caikit Text Embeddings

Typical Retrieval-Augmented Generation (RAG) Pattern flow



Orchestrate

IBM watsonx Orchestrate | Personal skills ▾ MB

A vaga solicitada teve a descrição criada, conforme abaixo, e já avisei às pessoas candidatas que se encaixam com o perfil por SMS!

Conteúdo gerado com sucesso!

A empresa é um banco de investimentos que está procurando um Gerente de Produto de Crédito - Cartão de Crédito e Débito para sua equipe em São Paulo. O candidato deve ter experiência em gestão de produtos de cartão de crédito e habilidades de comunicação. O Gerente de Produto será responsável por gerenciar o portfólio de cartões de crédito e débito, bem como por desenvolver e implementar estratégias para aumentar a participação de mercado e a satisfação do cliente.

Além disso, o Gerente de Produto será responsável por:

Tell me what you want to do

Add skills from the catalog →

Box 2 skills

Get all event invitees

Composite

GitHub 2 skills 7 skills

Geração d Conteúdo

<https://ibm.biz/wxo-brasil>



Explore Content

All Courses 129 Guided Projects 284

Sort ▾

Reset Filters

Results (85)

- Skills
- Tier
- Topic
- Difficulty
- Platform
- Language
- Offered By
- Industries
- Ratings



Beginner Course

Prompt Engineering for Everyone

Prompt Engineering, Generative AI, LLM, +1 more

9.72k+ Enrolled

5 Hours

★★★★★ (1.03k+)



Beginner Course

Introducing AI

Artificial Intelligence

6.34k+ Enrolled

1 Hour

★★★★★ (1.65k+)



Beginner Course

AI Ethics

Artificial Intelligence, Ethics, Bias, +2 more

2.29k+ Enrolled

1 Hour

★★★★★ (435)



Beginner Course

AI Concepts

Artificial Intelligence, Machine Learning, Deep...

1.73k+ Enrolled

1 Hour

★★★★★ (348)



Intermediate Guided Project

Build a Chatbot to Analyze PDF Documents Using LLM

Python, Generative AI, LLM, +2 more

570 Enrolled

60 Min

★★★★★ (58)



Beginner Guided Project

Build a Chatbot in less than an hour with watsonx!

Chatbots, Generative AI, Artificial Intelligenc...

600 Enrolled

1 Hour

★★★★★ (103)



Programa de cursos integrados Generative AI Fundamentals

Unlock and leverage the potential of generative AI. Learn how you can use the capabilities of generative AI to enhance your work and daily life.

Ministrado em Inglês



Instrutores: [IBM Skills Network Team](#) + Mais 2

Cadastre-se gratuitamente
Inicia em Nov 30

Auxílio financeiro disponível



Obrigado

