Potencializando Desenvolvimento de Software com o Modelo Open Source Granite-Code e watsonx

Alan Braz

AIware Latam 2024
https://link.pullrecast.dev/aiware

https://research.ibm.com/blog/granite-code-models-open-source

Agenda

- IBM Research Intro
- Plataforma watsonx
- IBM Granite Models
  - Granite Code Models
  - Granite for Function Calling

IBM Research

A community of 3,000

Hybrid Cloud

Artificial Intelligence

Quantum Computing

Semiconductors and Systems

Security and Cryptography

Physical Sciences

Mathematical Sciences

Computer Science

6 Nobel Laureates

10 Medals of Technology

5 National Medals of Science

6 Turing Awards


Dublin
Warrington
Zurich
Albany
Yorktown
Cambridge
Almaden
Haifa
Delhi
Tokyo
Shin-Kawasaki
Bangalore
Singapore
Nairobi
Rio de Janeiro
São Paulo
Johannesburg

Using AI technologies to strengthen **Indigenous Languages**


Paulo Cavalin

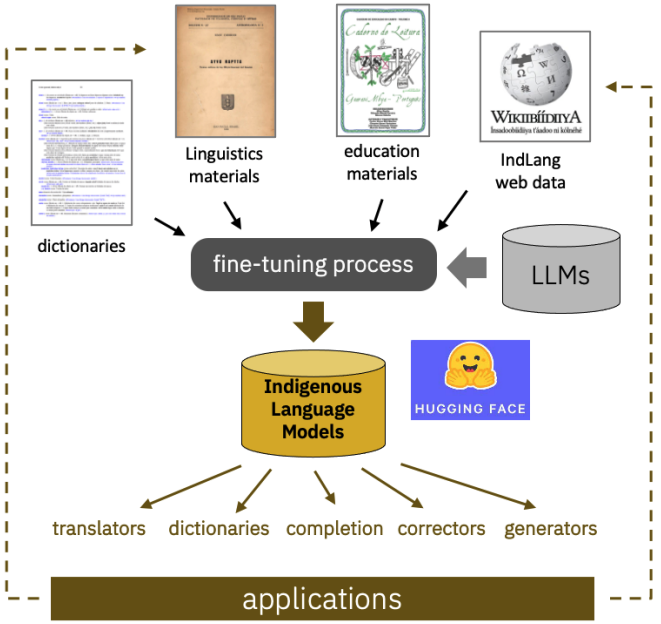

Pedro Domingues


Claudio Pinhanez


Julio Nogima

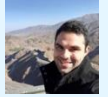
Technical strategy: Fine-Tuning LLMs with Linguistic Data to Create **Indigenous Language Models**

- To **deconstruct dictionaries** to generate high-quality lexical and translation datasets.
- To collect data from linguistic thesis, webpages, documents **to create monolingual datasets**.
- To use the high-quality datasets as guides to **expand with synthetic data** the monolingual and translation datasets.
- To **avoid the need of detailed linguistic knowledge** to create supporting tools by using a single FM.
- To **explore related languages** to enlarge training sets (multilingual approach).
- To make all **data, code, and models available in open-source** such as Hugging Face, under the **control of the Indigenous communities**.



Supporting the Discovery of Sustainable Materials

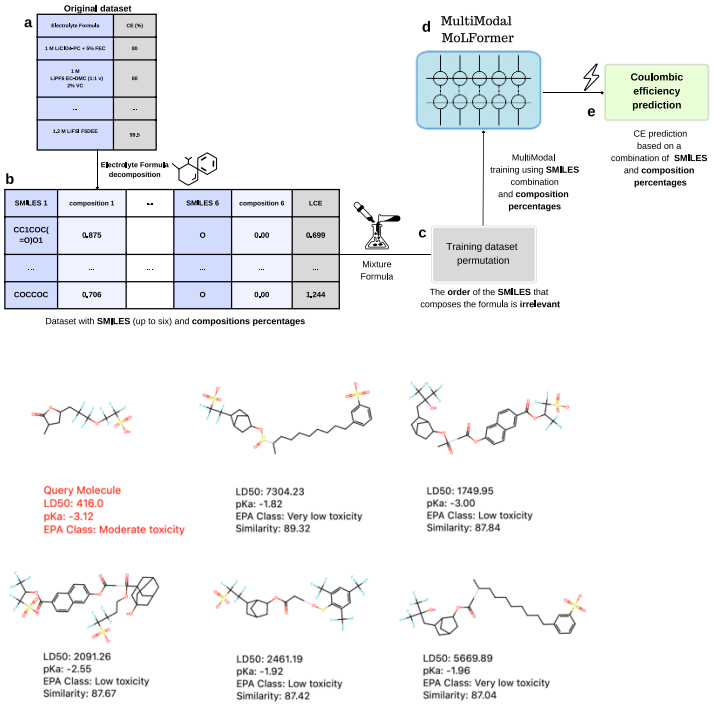
Technology Driven by the Use Cases


Eduardo Soares

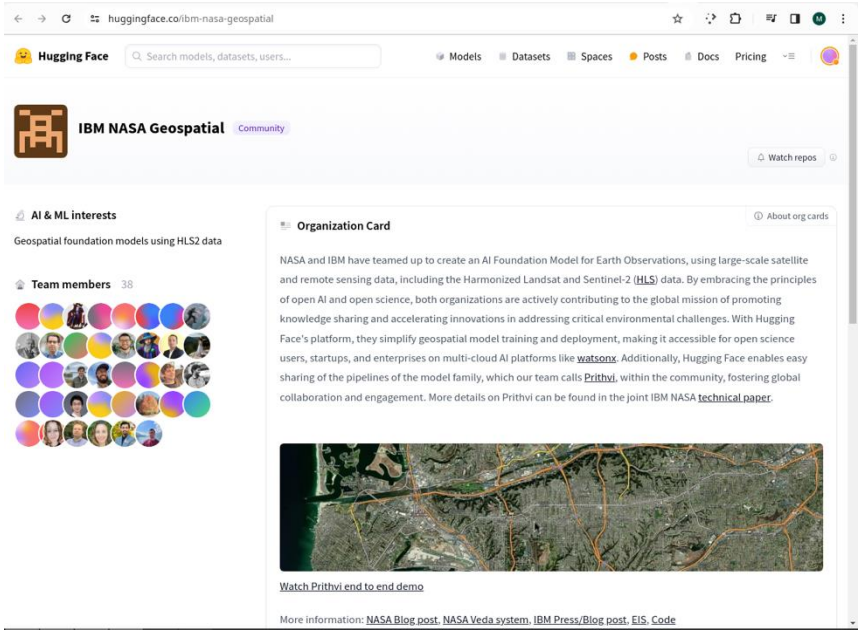

Emílio Vital Brazil

- Three challenging **Sustainable Materials** problems:
  - PFAS - A very useful material but toxic; how to replace it
  - Energy Storage - Find efficient materials that are safer
  - MoF - Discover new materials to clean pollution
- **Foundation Models** for properties prediction:
  - MOLFormer and RHIZOME – Molecule based models trained more than 1 Billion of Molecules and finetuned to tenths of downstream tasks related with the use cases
- **Related papers:**

Improving Molecular Properties Prediction Through Latent Space Fusion
Capturing Formulation Design of Battery Electrolytes with Chemical Large Language Model
Beyond Chemical Language: A Multimodal Approach to Enhance Molecular Property Prediction
A Framework for Toxic PFAS Replacement based on GFlowNet and Chemical Foundation Model



Foundation Models de IA para clima e tempo



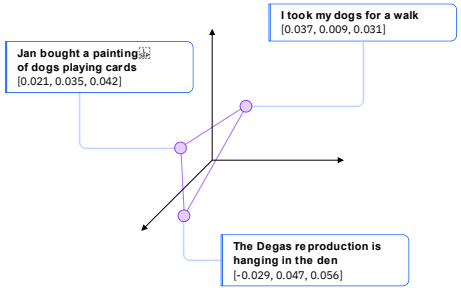
https://videos.netshow.me/t/2FqLo_jHbHM

watsonx Platform Engineering

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

Guardrails

Embedding



Performance na inferência

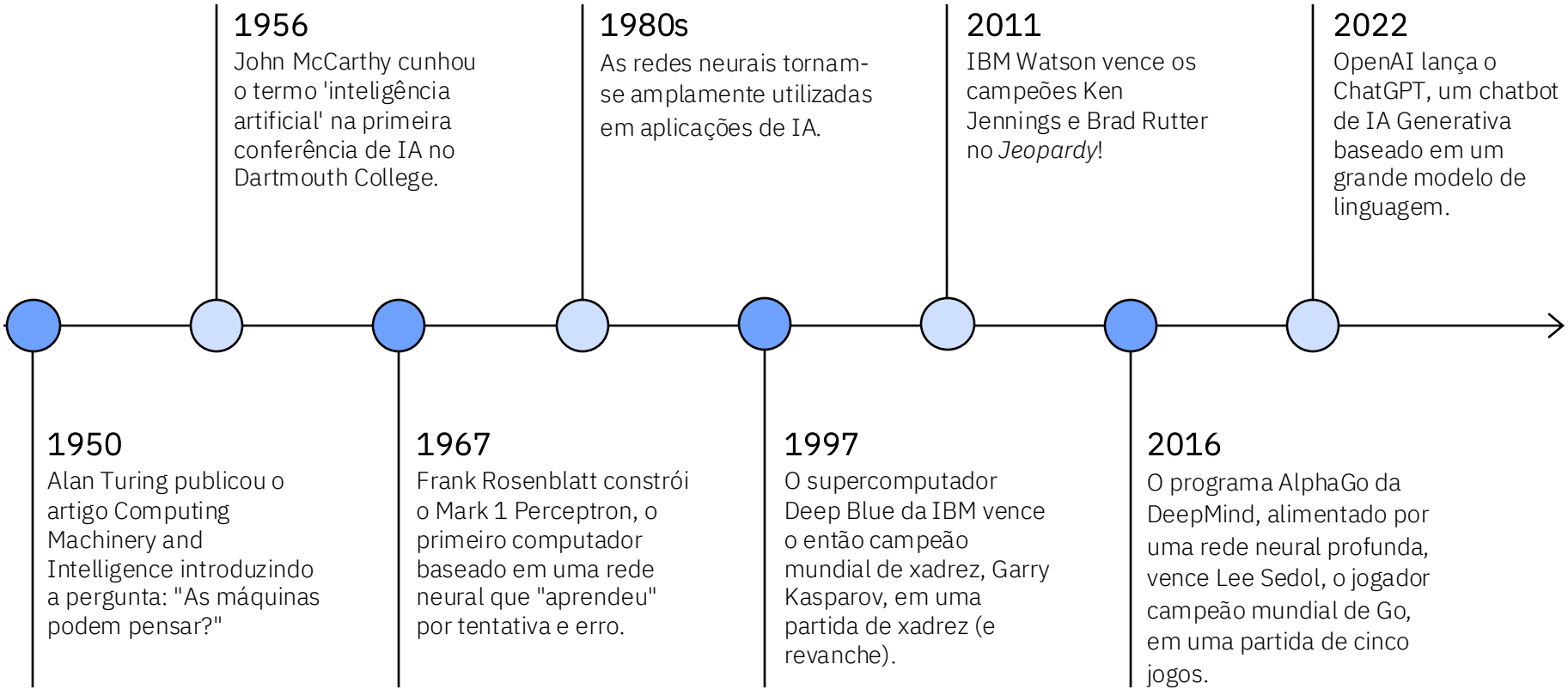
vLLM

TGI

Inferência em AIU

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

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

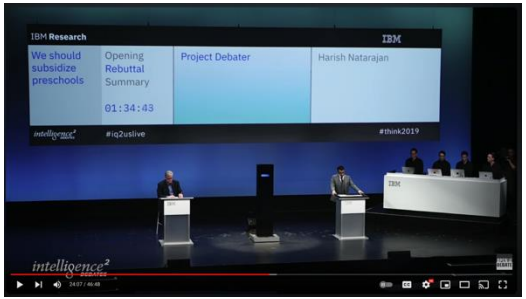
Em investida inédita no mundo, fragrâncias foram desenvolvidas em parceria com a IBM e chegam ao mercado no dia 27. Conheça em primeira mão a novidade

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

Watson 2011

Chef Watson 2017

Project Debater



IBM POV:
Quatro princípios fundamentais para adaptar a IA generativa para empresas

Aberta

→ Baseado nas melhores tecnologias de IA e nuvem disponíveis

→ Dar acesso à inovação da comunidade aberta e a múltiplos modelos

Focada

→ Projetado para casos de uso de negócios direcionados, que desbloqueiam um novo valor

→ Incluindo modelos selecionados que podem ser ajustados a dados proprietários e diretrizes da empresa

Confiável

→ Construído com IA e governança de dados, transparência e ética que suportam as crescentes demandas de conformidade regulatória

→ Fornecer orientação sobre modelos apropriados para alavancar para criar valor comercial real com confiança

Capacitante

→ Em uma plataforma para trazer seus próprios dados e modelos de IA que você ajusta, treina, implanta e controla

→ Execução em qualquer lugar, projetado para escala e adoção generalizada para realmente criar valor empresarial

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

Tecnologia e experiências de IA Generativa da IBM

AI assistants	Capacite os indivíduos a trabalhar sem conhecimento especializado em uma variedade de processos e aplicativos de negócios.	**watsonx** Code Assistant **watsonx** Assistant **watsonx** Orchestrate
SDKs and APIs	Incorpore a plataforma watsonx em assistentes e aplicativos de terceiros usando interfaces programáticas.	**Ecosystem integrations**
AI and data platform	Aproveite a IA generativa e o aprendizado de máquina, ajustados aos seus dados, com responsabilidade, transparência e explicabilidade.	**watsonx** watsonx.ai watsonx.governance watsonx.data **Foundation models** *IBM models* | Granite 3rd party models | *Meta, Mistral AI* | *Hugging Face*
Data services	Defina, organize, gerencie e forneça dados confiáveis para treinar e ajustar modelos de IA com serviços de malha de dados.	**Cloud Pak for Data** **watsonx** Discovery
Hybrid cloud AI tools	Construa sobre uma base consistente e escalável baseada em tecnologia de código aberto.	**Red Hat OpenShift AI** **RHEL AI** Granite Foundation Models InstructLab

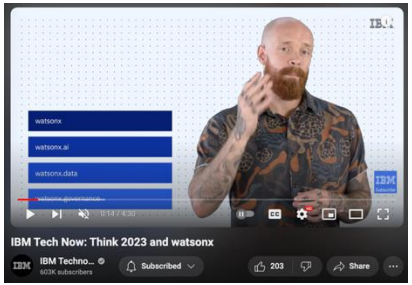
Consulting
Generative AI strategy, experience, technology, operations

Ecosystem
System Integrators, Software and SaaS partners, Public Cloud providers

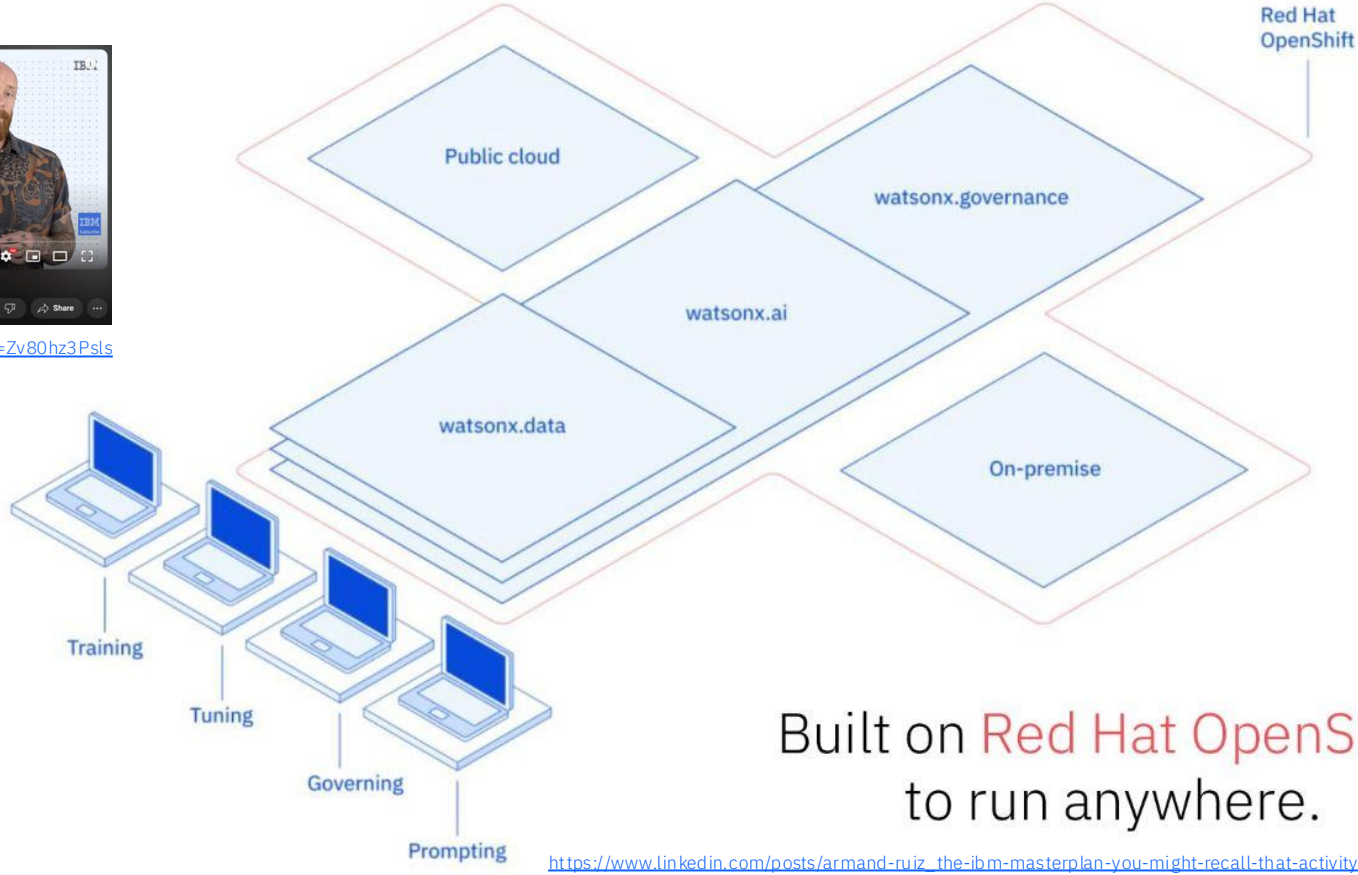
How to scale AI across the business

ibm.com/br-pt/watsonx

Putting AI to work on Hybrid Cloud



https://www.youtube.com/watch?v=Zv80hz3Psls



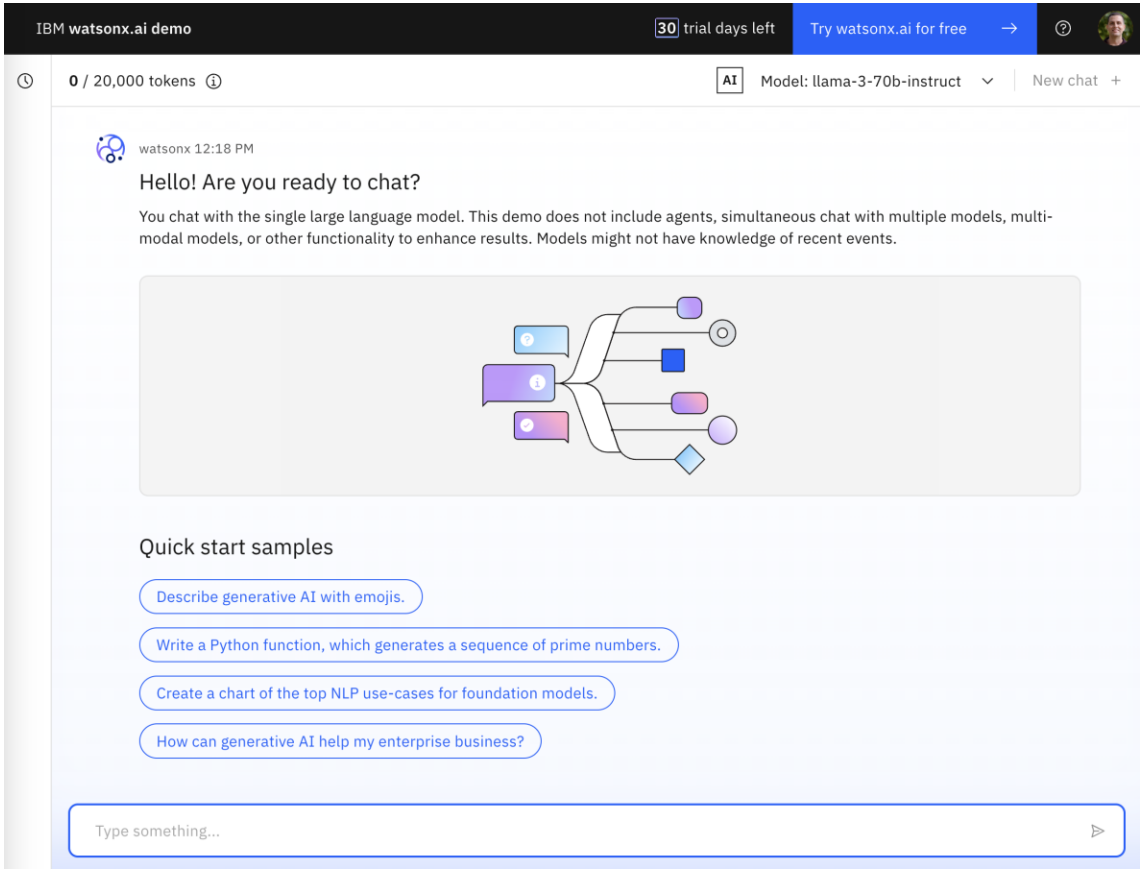
Built on Red Hat OpenShift.
to run anywhere.

https://www.linkedin.com/posts/armand-ruiz_the-ibm-masterplan-you-might-recall-that-activity-7134874545273937920-1

https://link.pullrecast.dev/wx-chat

Experimentar a nova interface de chat gratuíta: múltiplos modelos granite, llama, mixtral testar português

Trial: 50,000 tokens/mês 22 modelos, parâmetros, API, SDK, AutoAI



IBM Granite Models

Granite for Code

Trained on 116 programming languages, Granite code models (3b, 8b, 20b, 34b) are optimized for enterprise software development workflows. These models have a range of uses, from simple code completion to complex application modernization tasks and on-device memory constrained use cases.

Granite for Time Series

Granite Time Series is a family of lightweight, pre-trained models for time-series forecasting trained on a collection of datasets spanning a business a range of business and industrial application domains. We have optimized Granite Time Series to run efficiently across a range of hardware configurations, meaning you can start using them today with a laptop.

Granite for Language

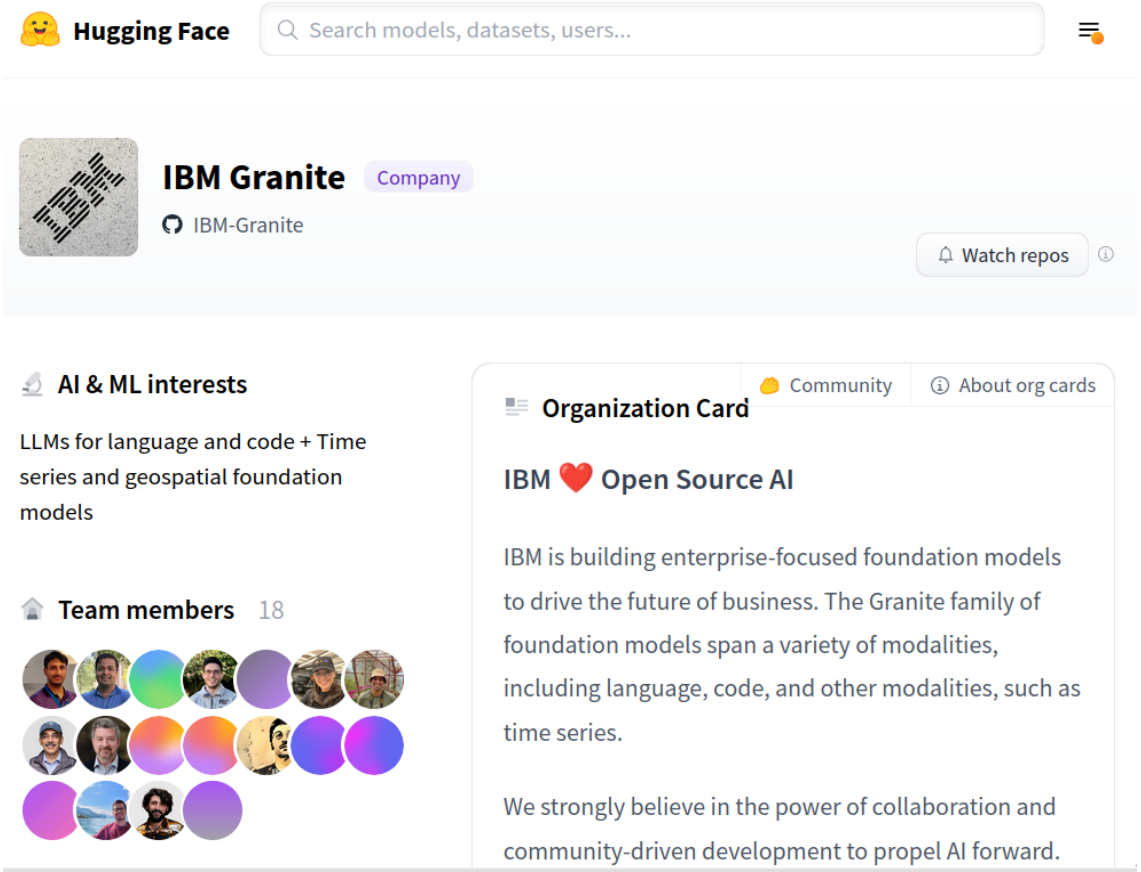
Granite language models (7b open-source, 13B English, 20b **multilingual**, 8b Japanese) demonstrate high accuracy and throughput at low latency, while consuming only a fraction of GPU resources.

Granite for GeoSpatial

NASA and IBM have teamed up to create an AI Foundation Model for Earth Observations, using large-scale satellite and remote sensing data, including the Harmonized Landsat and Sentinel-2 (HLS (link resides outside of ibm.com) data. By embracing the principles of open AI and open science, both organizations are actively contributing to the global mission of promoting knowledge sharing and accelerating innovations in addressing critical environmental challenges.

Open Source

https://huggingface.co/ibm-granite

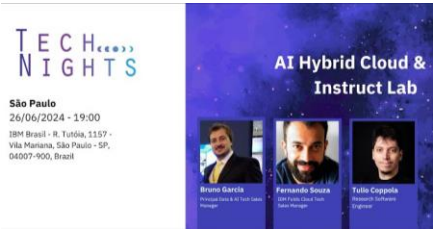


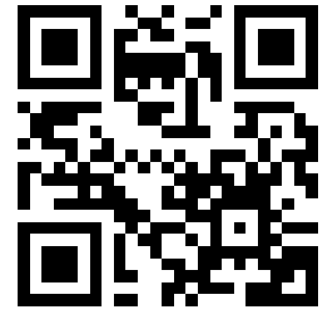
Open Source LLMs

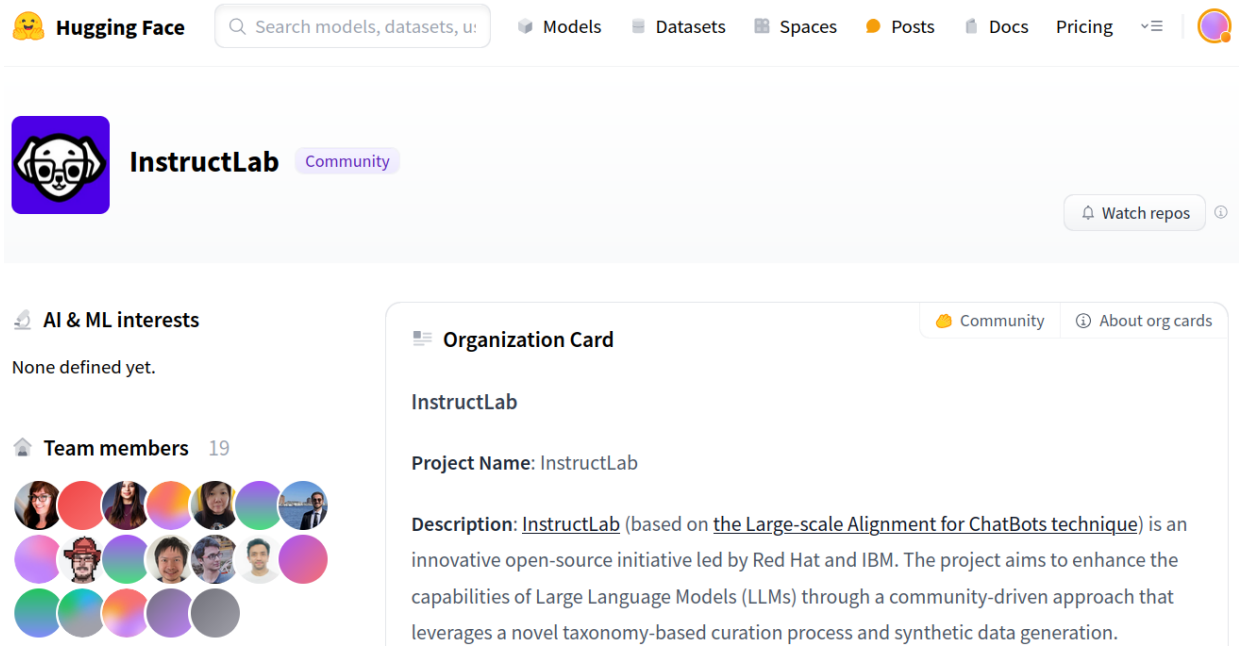
Acesso	Descrição	Exemplos
Fechado	• Acesso por API, sem acesso direto aos pesos do modelo ou ao código • Licença específica do provedor	• GPT4 • Claude 3
Open Weights	• Acesso aos pesos e código do modelo • Licença com restrições de uso	• LLama 3 • Gemma 2
Open Source	• Acesso aos pesos e código do modelo • Licença sem restrições de uso	• Granite • Mistral

InstructLab

https://youtu.be/ts0jejuSmkc?t=5054







IBM Granite Code Models

Granite Code Models: A Family of Open Foundation Models for Code Intelligence: https://arxiv.org/abs/2405.04324

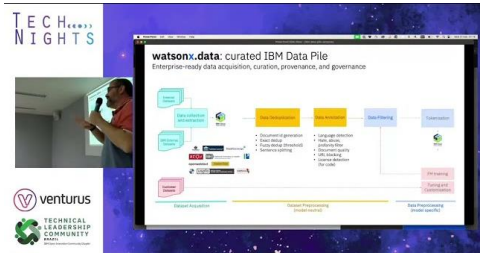
ibm-granite - granite-code-models collection

- Modelos treinados com datasets de código
- 4 tamanhos disponíveis: 3b, 8b, 20b e 34b
- 2 Variantes de treinamento para cada tamanho: base e instruct
- 2 Variantes de tamanho de contexto

IBM Granite Code Models

	3b	8b	20b	34b
Architecture	Llama	Llama	GPTBigCode	GPTBigCode
Context Length	2048	4096	8192	8192
Hidden Size	2560	4096	6144	6144
Attention Heads	32	32	48	48
Key-Value Heads	32 (MHA)	8 (GQA)	1 (MQA)	1 (MQA)
Layers	32	36	52	88
Vocab. Size	49152	49152	49152	49152
Pos. Embeddings	Relative	Relative	Absolute	Absolute

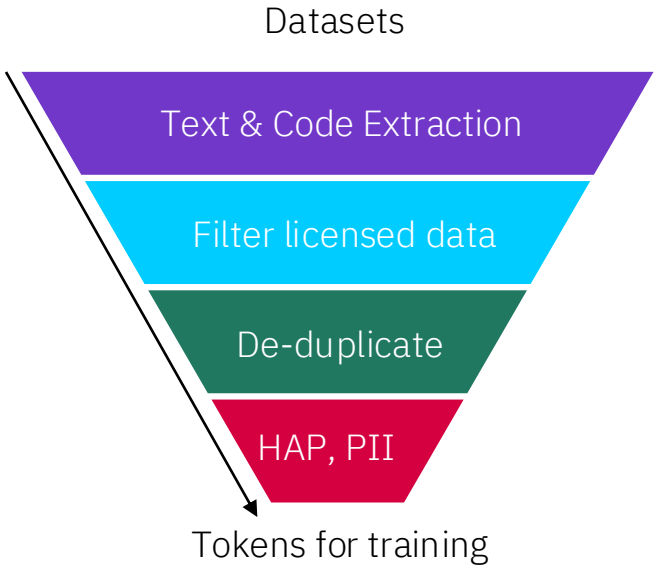
Pre-Training: granite-Xb-base



https://link.pullrecast.dev/tn12

Objetivos de treinamento:

- Causal language modeling
- "Fill in the middle"



Duas fases de treinamento:

- **Fase 1**:
  - 3b e 8b -> 4 trilhões de tokens de código
  - 20b -> 3 trilhões de tokens de código
  - 34b -> 1.4 trilhões de tokens após o upscaling
- **Fase 2**: Código e linguagem natural: 500 bilhões de tokens

Dataset:

- 116 Linguagens de programação
- Datasets de linguagem natural voltados para matemática e código
- Filtros de HAP, PII e Malware
- DataPrepKit: https://github.com/IBM/data-prep-kit

Instruction Tuning: granite-Xb-instruct

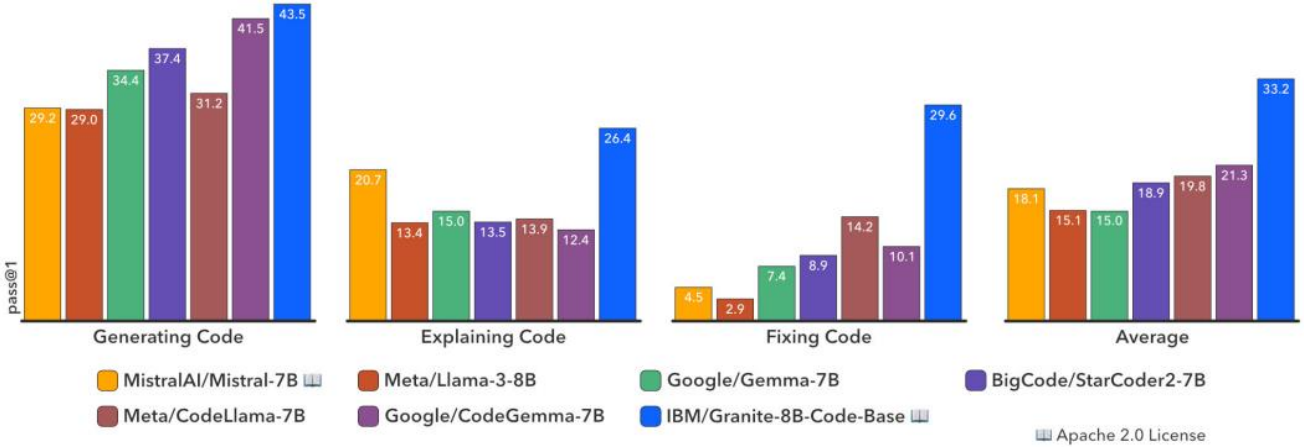
- Geração de código em 18 linguagens
- Explicação de código
- Conserto de código
- Edição de código
- Tradução de código
- Raciocínio, Compreensão e execução
- Raciocínio Matemático
- Chamada de funções

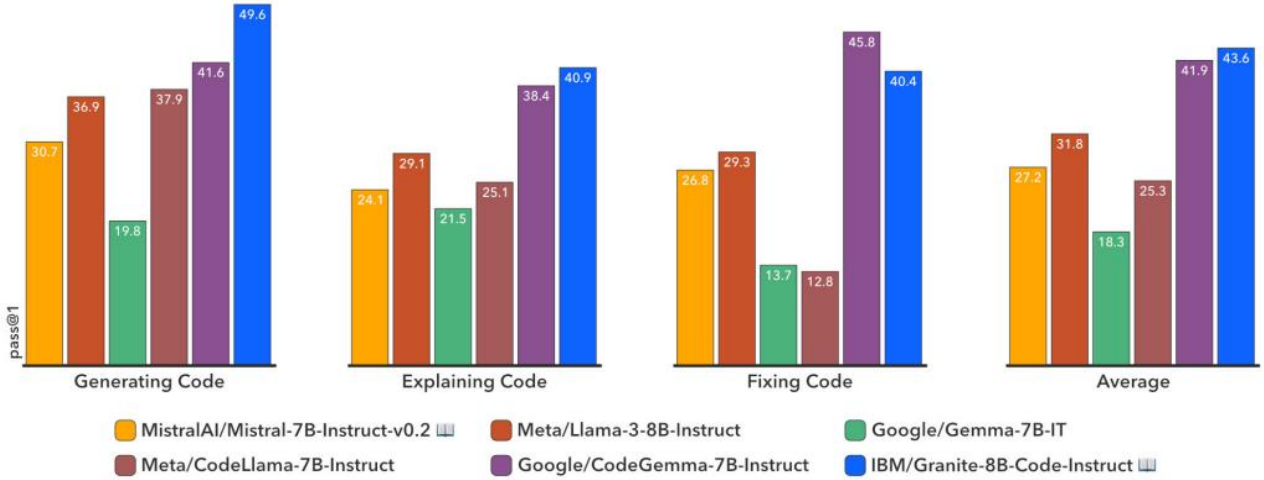
Resultados

Comparação Granite-8B-Code (Base/Instruct) com outros LLMs abertos de tamanho simliar no benchmark HumanEvalPack

Resultados na data da publicação, 7 de Maio 2024

Mais detalhes: https://arxiv.org/abs/2405.04324





Long Context - 128k

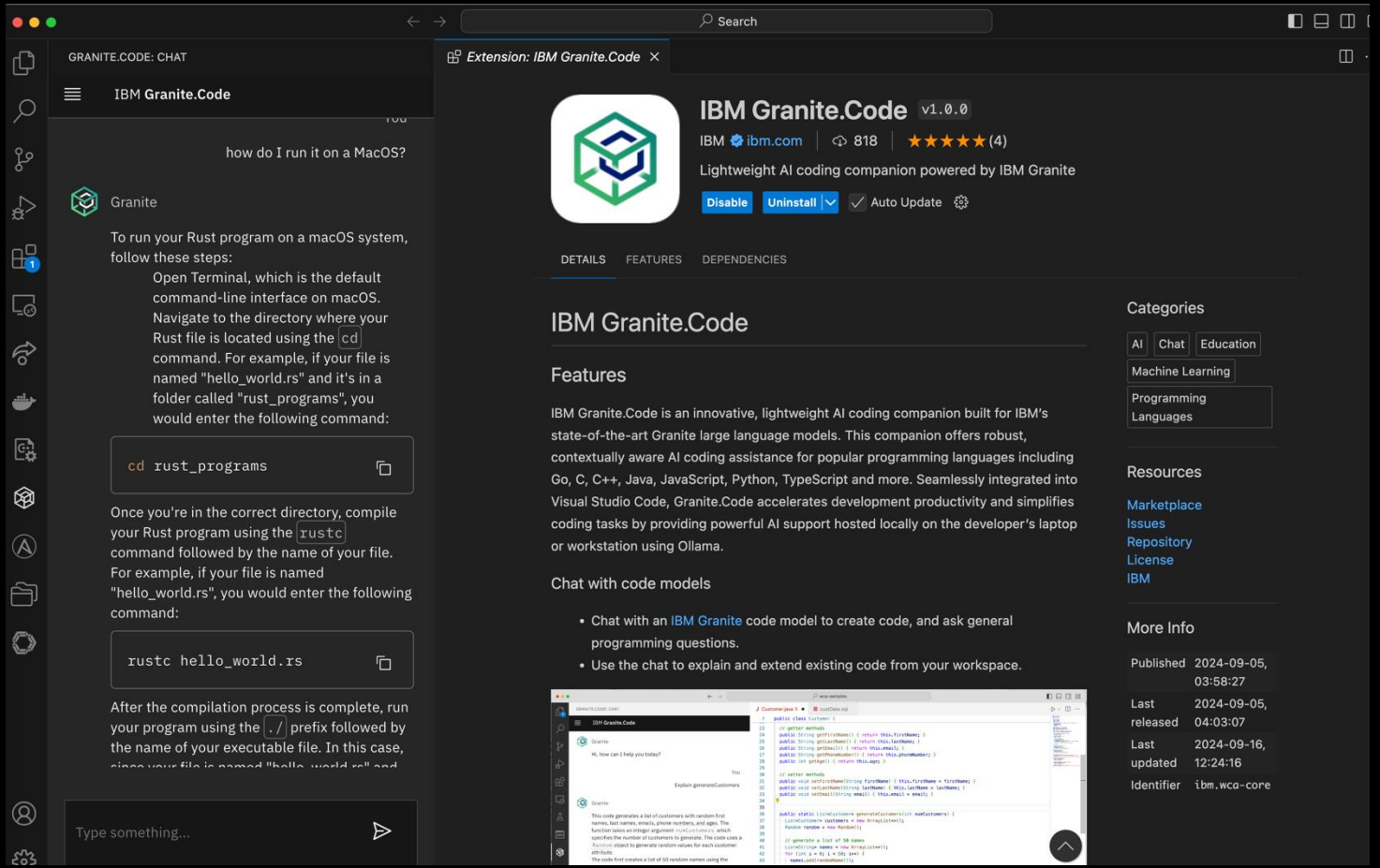
Data Engineering for Scaling Language Models to 128K Context: https://arxiv.org/abs/2402.10171

Scaling Granite Code Models to 128K Context: https://arxiv.org/abs/2407.13739

Long Context Understanding using Self-Generated Synthetic Data: https://openreview.net/forum?id=EciNEiW1uq

- Insight: Modelos treinados em contextos curtos como 4k podem ser treinados para contextos longos a uma fração do custo e número de tokens.
- Para atingir qualidade em 80K -> 500M de tokens
- Qualidade até 128k 5b tokens
- Treinamento de llama-2 7b leva 5 dias em 8 GPUS A100
- Parâmetros do embedding posicional relativo são ajustados para contextos longos

Demo



Alternativa: Build a local AI co-pilot using IBM Granite Code, Ollama, and Continue: https://allthingsopen.org/articles/build-a-local-ai-co-pilot

Granite-20B-FunctionCalling

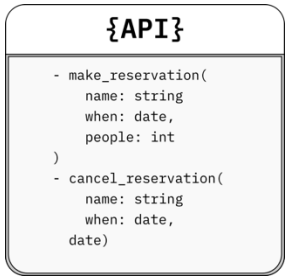
Granite-Function Calling Model: Introducing Function Calling Abilities via Multi-task Learning of Granular Tasks
https://arxiv.org/abs/2407.00121

https://huggingface.co/ibm-granite/granite-20b-functioncalling

Fine-tuning do modelo Granite-20b-Code-Instruct

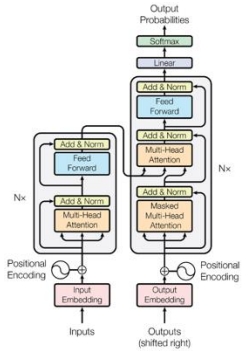
Características	
Licença	Apache 2.0
Publicação	Julho 2024
Arquitetura	GPTBigCodeForCausalLM
Vocabulário	49152
N. Heads	48
Dim. Embedding	6144
Camadas	52

Granite-20B-FunctionCalling

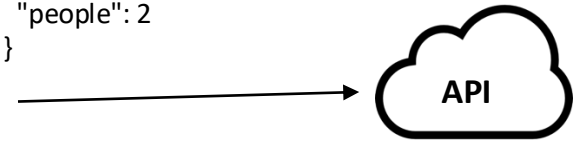


"Reserve uma mesa"

"Reserva feita"



POST: http://restaurant.com/reservation {
"when": "30/08/2024",
"name": "John",
"people": 2
}



https://arxiv.org/abs/2407.00121
https://huggingface.co/ibm-granite/granite-20b-functioncalling

Obrigado!

alanbraz@br.ibm.com