

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

Physical
Sciences

Artificial
Intelligence

Mathematical
Sciences

Quantum
Computing

Computer
Science

Semiconductors
and Systems

Security and
Cryptography



6 Nobel Laureates



10 Medals of Technology



5 National Medals of Science



6 Turing Awards



Using AI technologies to strengthen Indigenous Languages



Paulo
Cavalin



Pedro
Domingues



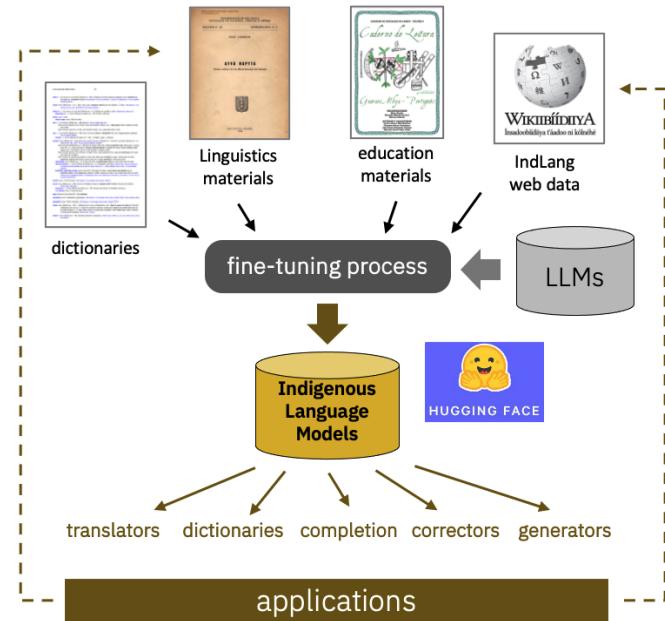
Claudio
Pinhanez



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

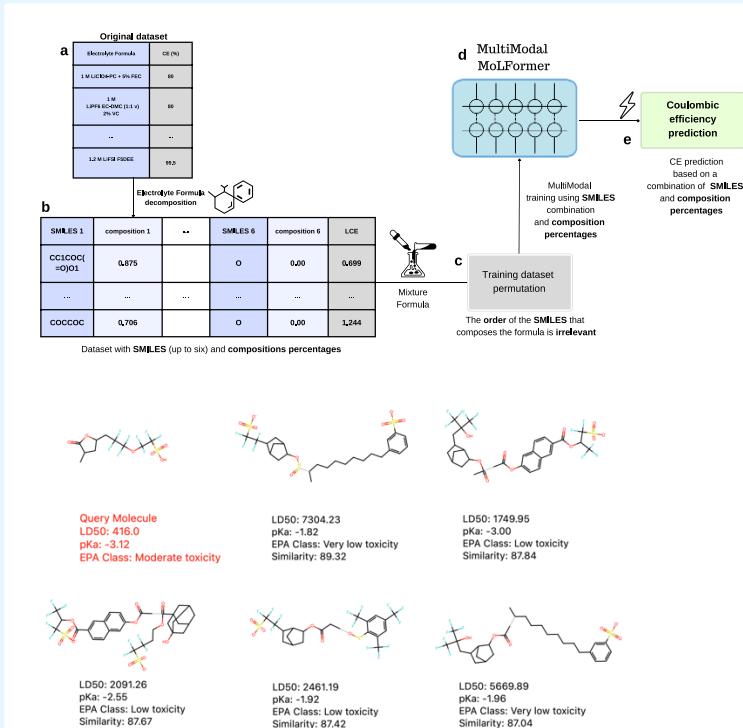
- Three challenging Sustainable Materials problems:
PFAS - A very useful material but toxic; how to replace it
Energy Storage - Find efficient materials that are safer
MoF - Discover new materials to clean pollution
- Foundation Models for properties prediction:
MOLFormer and **RHIZOME** – Molecule based 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

Bluetalks:

Foundation Models de IA para clima e tempo

Speakers
Daniel Salles Civitarese
Research Scientist

Data
7 de agosto de 2024

Horário
18h - 19h

IBM Research Brasil

Apoio: **inovabra**



huggingface.co/ibm-nasa-geospatial

Hugging Face Search models, datasets, users... Models Datasets Spaces Posts Docs Pricing

IBM NASA Geospatial Community

Watch repos

AI & ML interests

Geospatial foundation models using HLS2 data

Team members 38



Organization Card

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) 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. With Hugging Face's platform, they simplify geospatial model training and deployment, making it accessible for open science users, startups, and enterprises on multi-cloud AI platforms like Watson. Additionally, Hugging Face enables easy sharing of the pipelines of the model family, which our team calls Prithvi, within the community, fostering global collaboration and engagement. More details on Prithvi can be found in the joint IBM NASA technical paper.



Watch Prithvi end to end demo

More information: [NASA Blog post](#), [NASA Veda system](#), [IBM Press/Blog post](#), [EIS Code](#)

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



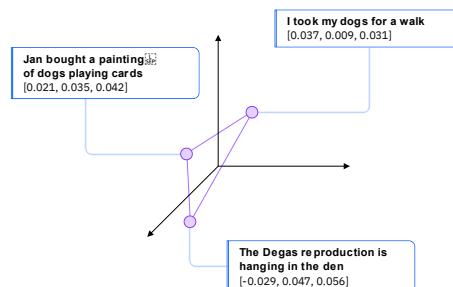
watsonx Platform Engineering

- 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

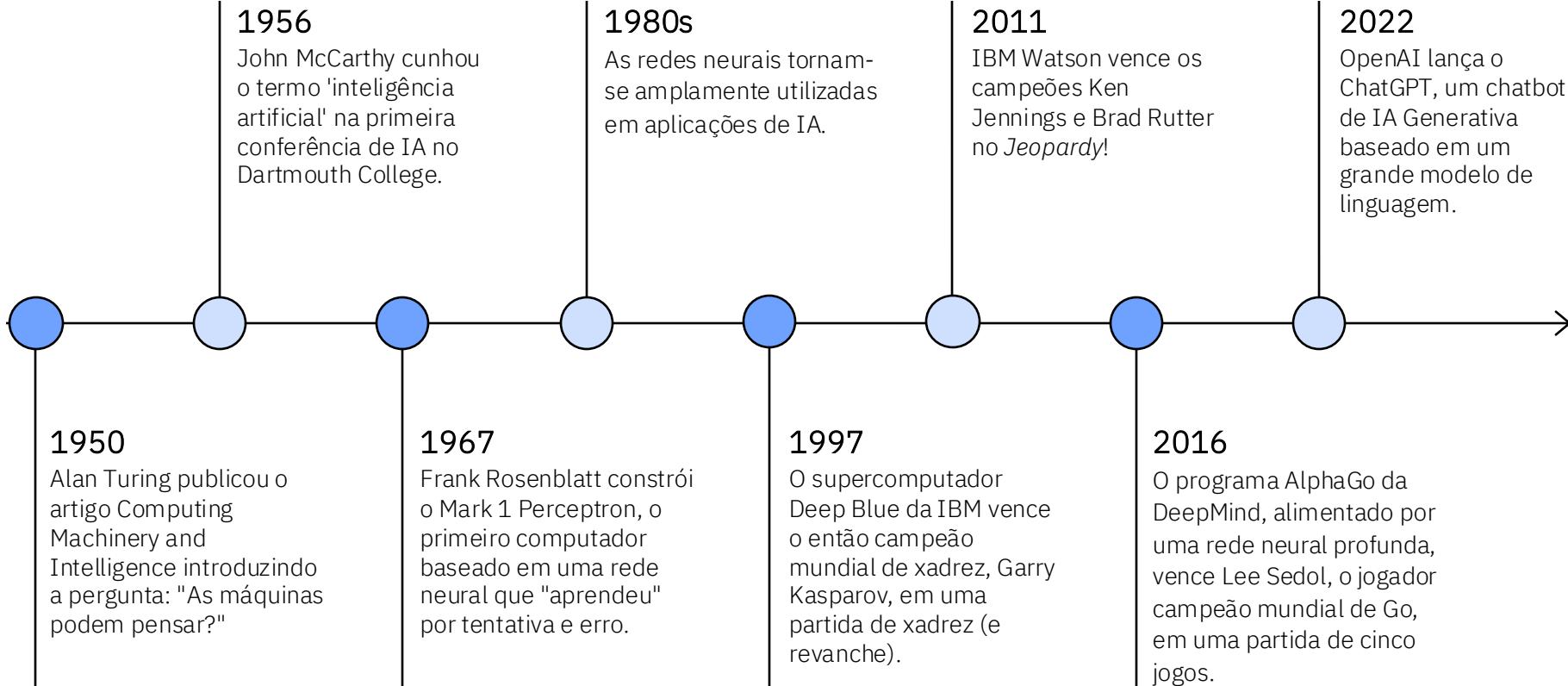
LLM
TGI

Inferência em AIU



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

Marcos da IA



Deep Blue 1997



Watson 2011



Ei! 2013/2014



Chef Watson 2017



2019

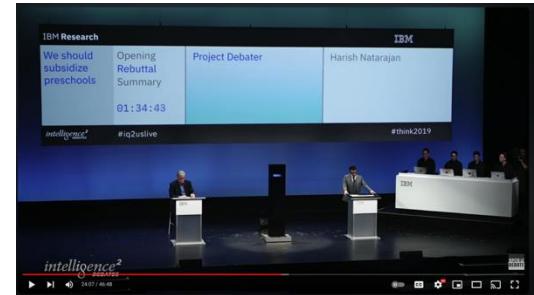
O Boticário lança 1ºs perfumes feitos com ajuda de inteligência artificial

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



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

Project Debater



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.

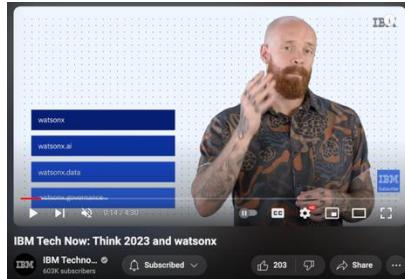
Red Hat OpenShift AI

RHEL AI
Granite Foundation Models
InstructLab

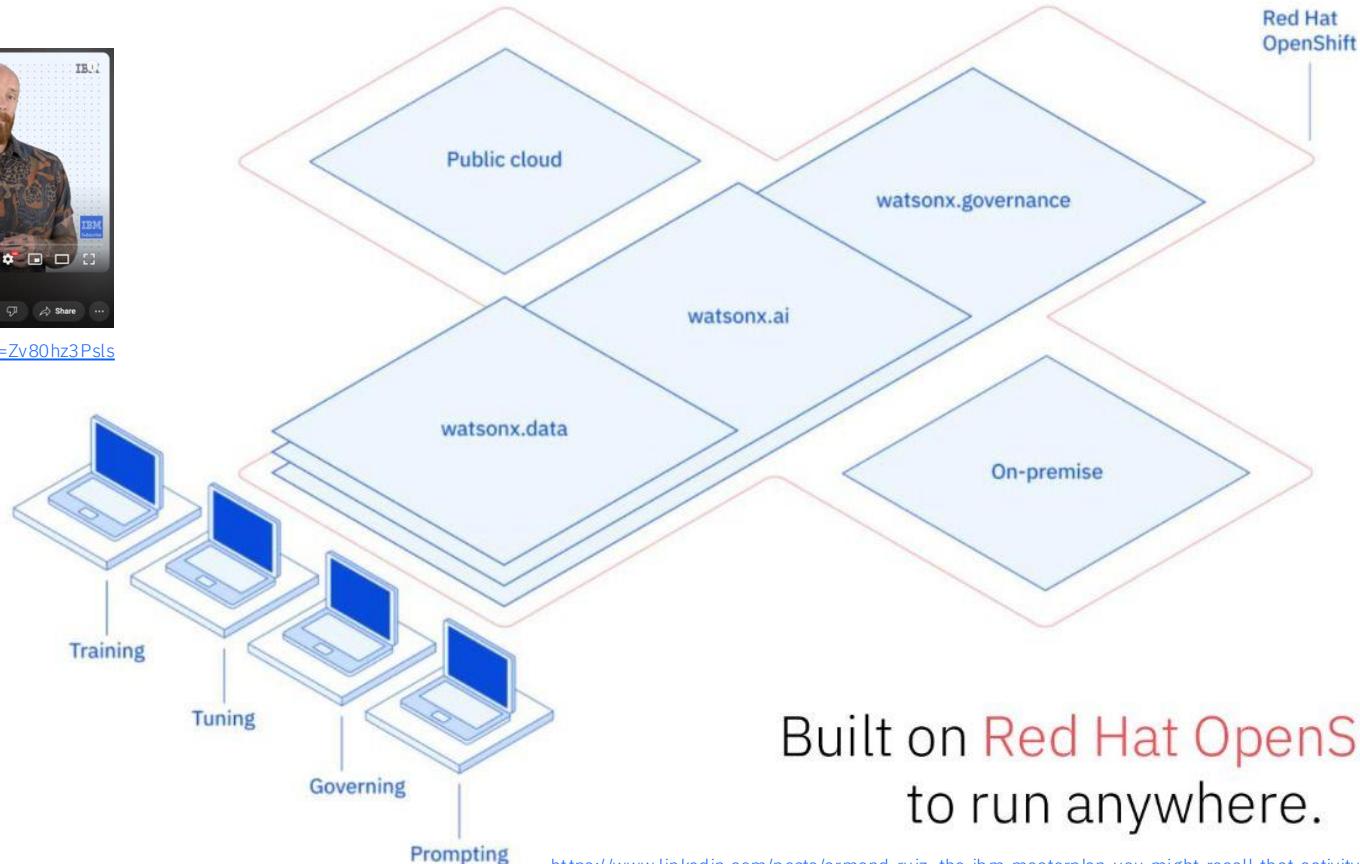
Consulting
Generative AI strategy, experience, technology, operations

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

Putting AI to work on Hybrid Cloud



<https://www.youtube.com/watch?v=Zv80hz3PsIs>



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

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

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

IBM watsonx.ai demo

30 trial days left Try watsonx.ai for free → ⚭ [Profile]

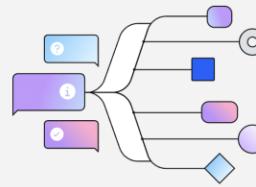
0 / 20,000 tokens ⓘ

AI Model: llama-3-70b-instruct ⓘ New chat +

watsonx 12:18 PM

Hello! Are you ready to chat?

You chat with the single large language model. This demo does not include agents, simultaneous chat with multiple models, multi-modal models, or other functionality to enhance results. Models might not have knowledge of recent events.



Quick start samples

- Describe generative AI with emojis.
- Write a Python function, which generates a sequence of prime numbers.
- Create a chart of the top NLP use-cases for foundation models.
- How can generative AI help my enterprise business?

Type something... ➤



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>

The screenshot shows the Hugging Face platform interface for the organization "IBM Granite". At the top, there's a search bar with the placeholder "Search models, datasets, users..." and a menu icon. Below the search bar, the organization's logo (a stylized "X" made of diagonal lines) and name "IBM Granite" are displayed, along with a "Company" badge. A user profile icon for "IBM-Granite" is shown next to the organization name. On the right, there's a "Watch repos" button and a small info icon. The main content area includes sections for "AI & ML interests" (describing LLMs for language and code + Time series and geospatial foundation models), "Team members" (listing 18 individuals with circular profile pictures), and an "Organization Card" (highlighting IBM's focus on Open Source AI and its mission to build enterprise-focused foundation models). The "Organization Card" also includes a "Community" tab and an "About org cards" link.

AI & ML interests
LLMs for language and code + Time series and geospatial foundation models

Team members 18

Organization Card

IBM ❤️ Open Source AI

IBM is building enterprise-focused foundation models to drive the future of business. The Granite family of foundation models span a variety of modalities, including language, code, and other modalities, such as time series.

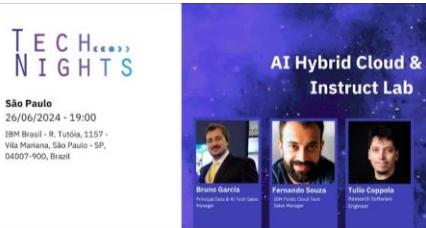
We strongly believe in the power of collaboration and community-driven development to propel AI forward.

Open Source LLMs

Acesso	Descrição	Exemplos
Fechado	<ul style="list-style-type: none">• Acesso por API, sem acesso direto aos pesos do modelo ou ao código• Licença específica do provedor	<ul style="list-style-type: none">• GPT4• Claude 3
Open Weights	<ul style="list-style-type: none">• Acesso aos pesos e código do modelo• Licença com restrições de uso	<ul style="list-style-type: none">• LLama 3• Gemma 2
Open Source	<ul style="list-style-type: none">• Acesso aos pesos e código do modelo• Licença sem restrições de uso	<ul style="list-style-type: none">• Granite• Mistral

InstructLab

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



Hugging Face Search models, datasets, u: Models Datasets Spaces Posts Docs Pricing

InstructLab Community

Watch repos

AI & ML interests

None defined yet.

Team members 19

Organization Card

Community About org cards

InstructLab

Project Name: InstructLab

Description: InstructLab (based on the Large-scale Alignment for ChatBots technique) is an innovative open-source initiative led by Red Hat and IBM. The project aims to enhance the capabilities of Large Language Models (LLMs) through a community-driven approach that leverages a novel taxonomy-based curation process and synthetic data generation. InstructLab provides tools for users to engage with and improve LLMs, contributing skills and knowledge to the project's taxonomy repository.

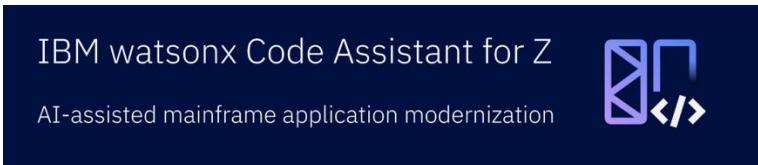
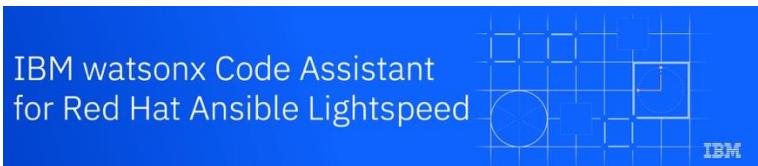
IBM Granite Code Models



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

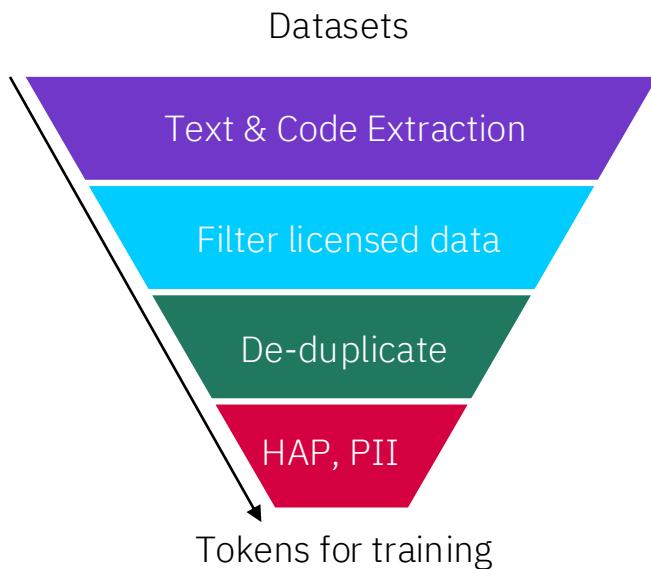
IBM Granite Code Models

	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

Objetivos de treinamento:

- Causal language modeling
- "Fill in the middle"

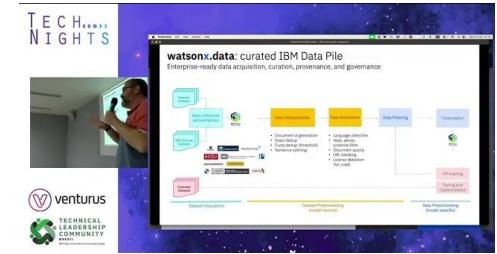


Duas fases de treinamento:

- **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>



<https://link.pullrecast.dev/tn12>

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

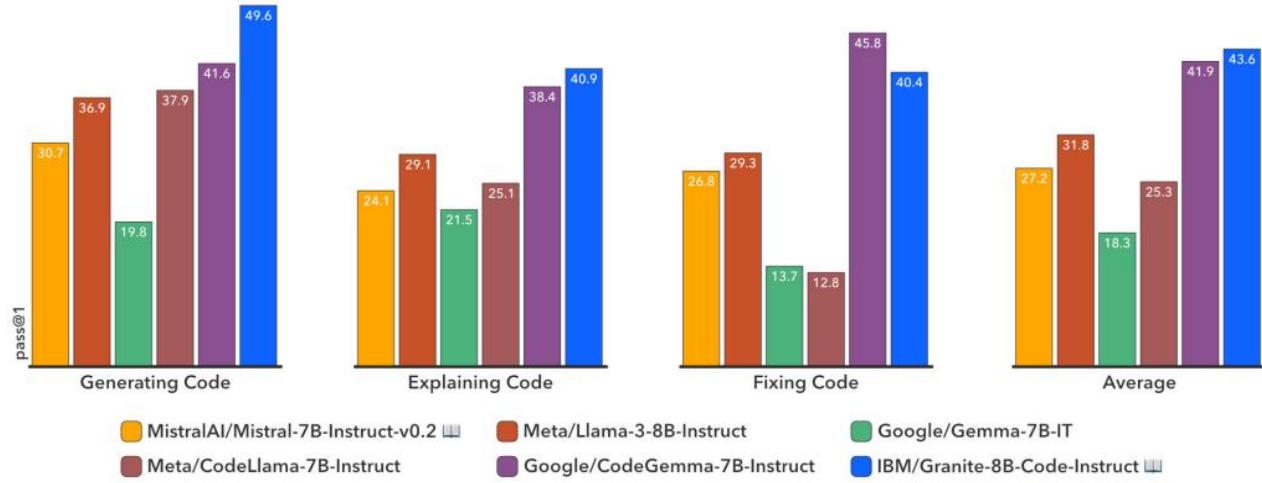
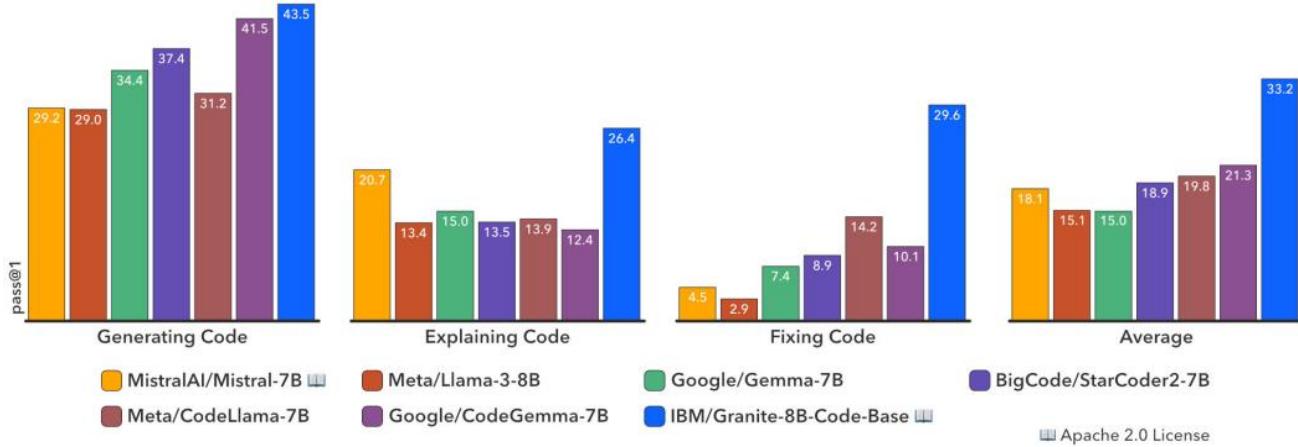
Resultados

Comparação Granite-8B-Code (Base/Instruct) com outros LLMs abertos de tamanho similar 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



The screenshot shows the Visual Studio Code interface with the "IBM Granite.Code" extension installed. The sidebar has a tree view with "GRANITE.CODE: CHAT" selected. The main panel shows a chat window with the message "how do I run it on a MacOS?". Below it, there's a section titled "Granite" with instructions: "To run your Rust program on a macOS system, follow these steps: Open Terminal, which is the default command-line interface on macOS. Navigate to the directory where your Rust file is located using the `cd` command. For example, if your file is named "hello_world.rs" and it's in a folder called "rust_programs", you would enter the following command:" followed by a terminal input box containing `cd rust_programs`. Further down, another section says "Once you're in the correct directory, compile your Rust program using the `rustc` command followed by the name of your file. For example, if your file is named "hello_world.rs", you would enter the following command:" followed by a terminal input box containing `rustc hello_world.rs`. At the bottom, there's a search bar with "Type something..." and a send icon.

The screenshot shows the IBM Marketplace page for the "IBM Granite.Code" extension. The extension icon is a green cube with a white "G" on it. The title is "IBM Granite.Code v1.0.0" with a link to "ibm.com". It has a rating of 4 stars from 4 reviews. The description is "Lightweight AI coding companion powered by IBM Granite". There are "Disable", "Uninstall", and "Auto Update" buttons. Below the extension card, there are sections for "DETAILS", "FEATURES", and "DEPENDENCIES". The "FEATURES" section contains the text: "IBM Granite.Code is an innovative, lightweight AI coding companion built for IBM's state-of-the-art Granite large language models. This companion offers robust, contextually aware AI coding assistance for popular programming languages including Go, C, C++, Java, JavaScript, Python, TypeScript and more. Seamlessly integrated into Visual Studio Code, Granite.Code accelerates development productivity and simplifies coding tasks by providing powerful AI support hosted locally on the developer's laptop or workstation using Ollama." The "Chat with code models" section lists two bullet points: "Chat with an IBM Granite code model to create code, and ask general programming questions." and "Use the chat to explain and extend existing code from your workspace." To the right, there are sections for "Categories" (AI, Chat, Education, Machine Learning, Programming Languages), "Resources" (Marketplace, Issues, Repository, License, IBM), and "More Info" (Published: 2024-09-05, 03:58:27, Last released: 04:03:07, Last updated: 2024-09-16, Identifier: ibm.wca-core). At the bottom, there's a preview image of the extension in action within VS Code.

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>

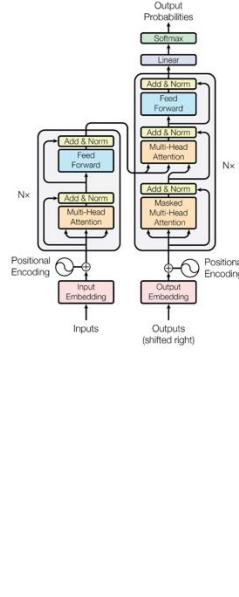
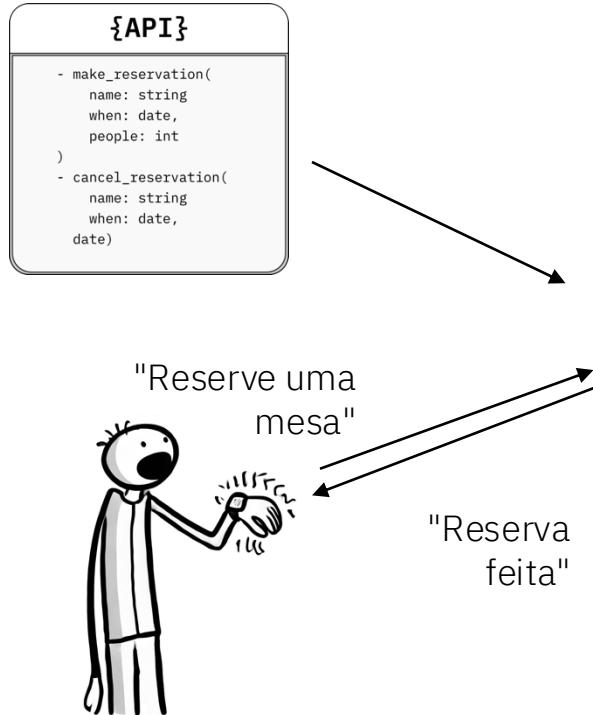


<https://huggingface.co/ibm-granite/granite-20b-functioncalling>

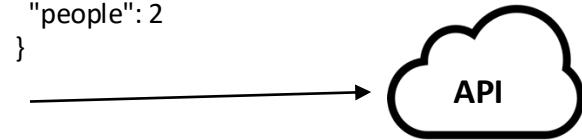
Fine-tuning do modelo Granite-20b-Code-Instruct

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

Granite-20B-FunctionCalling



```
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

© 2024 International Business Machines Corporation

IBM and the IBM logo are trademarks of IBM Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on ibm.com/trademark.

This document is current as of the initial date of publication and may be changed by IBM at any time.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER EXPRESS OR IMPLIED. IN NO EVENT SHALL IBM BE LIABLE FOR ANY DAMAGE ARISING FROM THE USE OF THIS INFORMATION, INCLUDING BUT NOT LIMITED TO, LOSS OF DATA, BUSINESS INTERRUPTION, LOSS OF PROFIT OR LOSS OF OPPORTUNITY.

Client examples are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Not all offerings are available in every country in which IBM operates.

It is the user's responsibility to evaluate and verify the operation of any other products or programs with IBM products and programs.

The client is responsible for ensuring compliance with laws and regulations applicable to it. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the client is in compliance with any law or regulation.