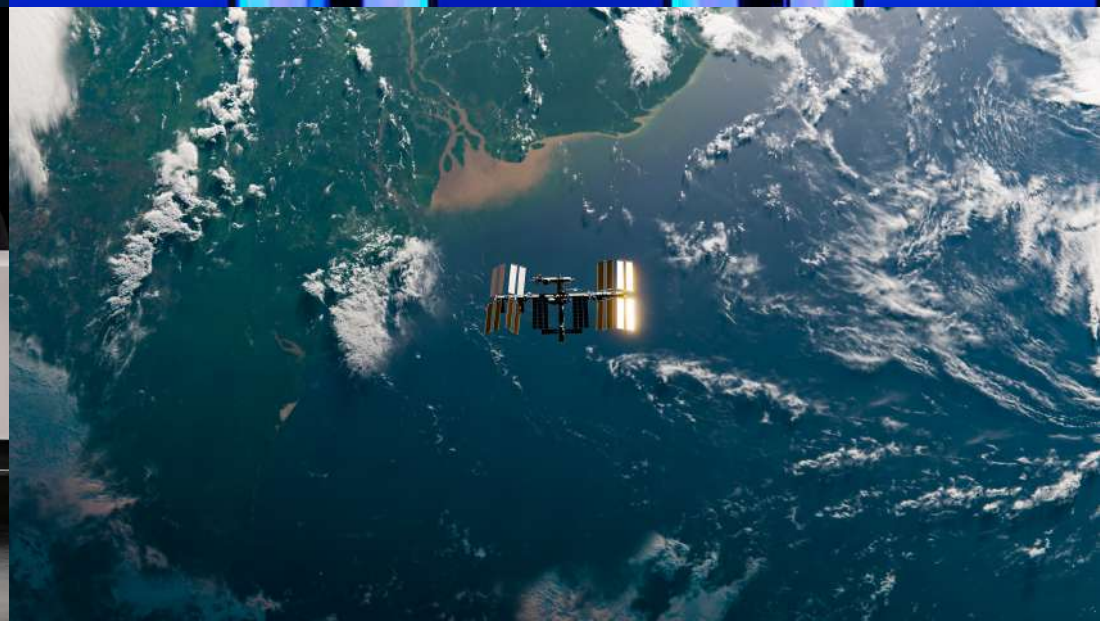
La Inteligencia Artificial para un Futuro Sostenible

Narciso Lema
CTO & Technical Community Leader IBM
Colombia, Perú, Ecuador, Venezuela & LCR
https://www.linkedin.com/in/narcisolema/


the rematch


THINK
$300,000
$1,000,000
$200,000
KEN
WATSON
BRAD


#think2019
IBM
IBM



El Futuro de la Inteligencia Artificial Empresarial

LLMs
Escalabilidad

Open
Colaboración

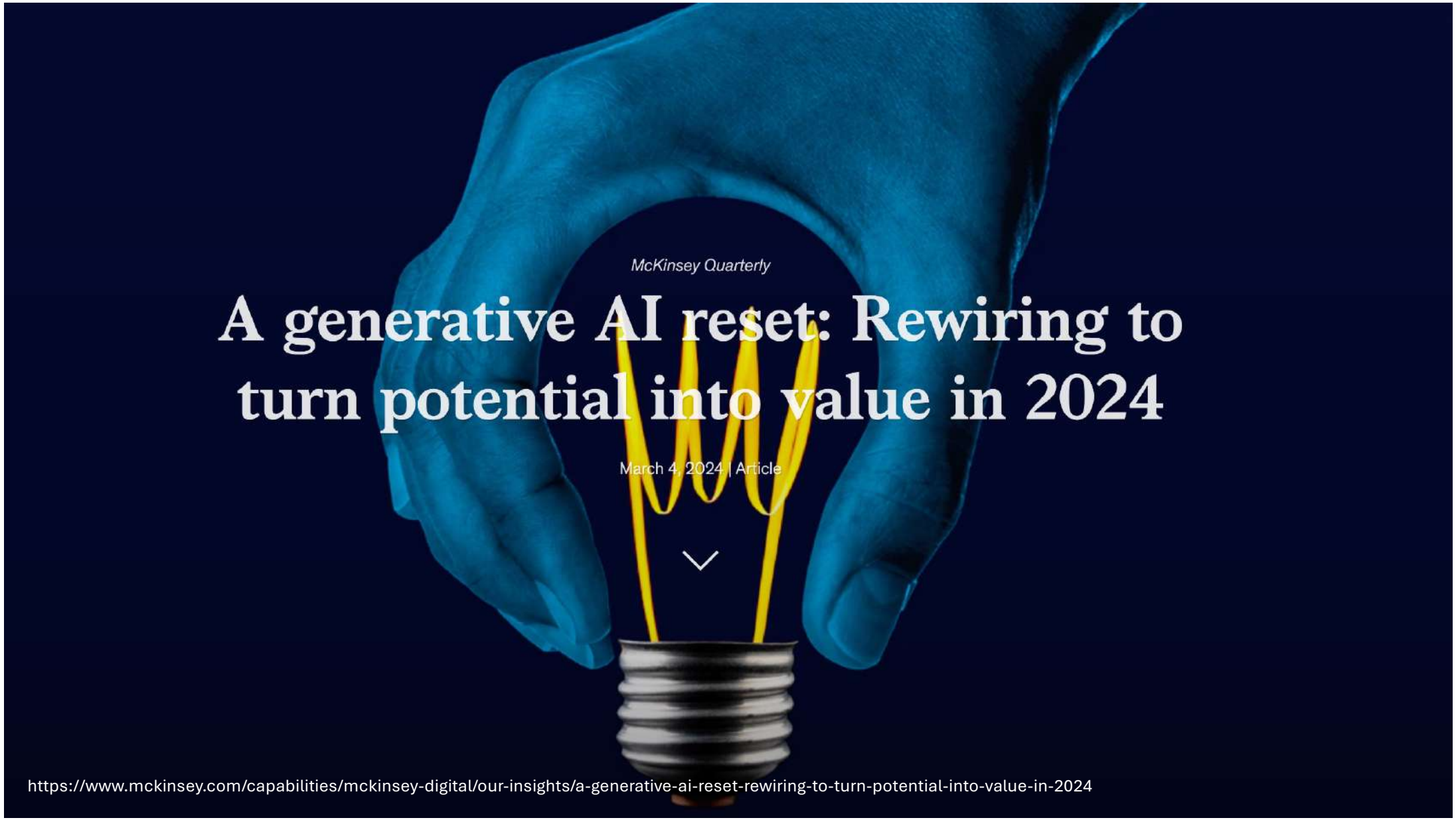
Quantum
Límites

El Futuro de la Inteligencia Artificial Empresarial

LLMs
Escalabilidad

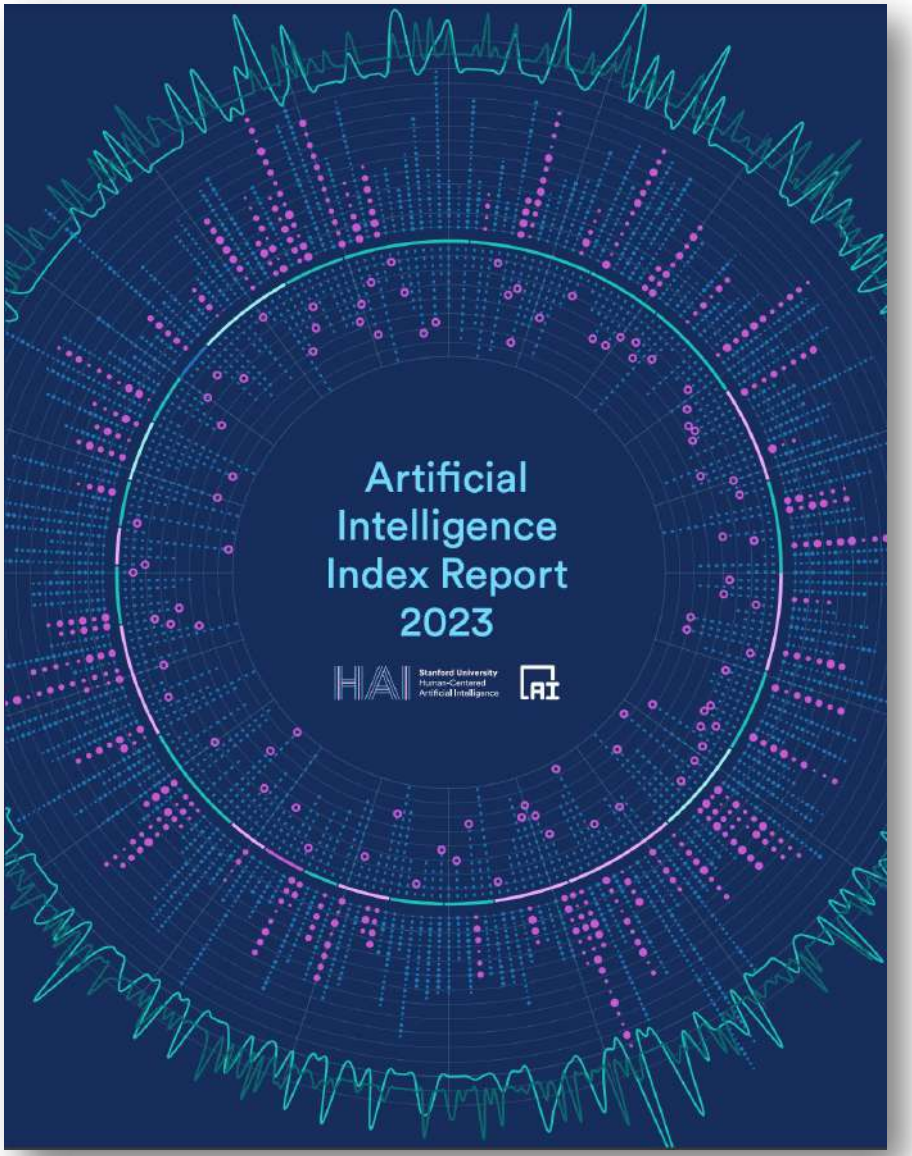
Open
Colaboración

Quantum
Límites



https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/a-generative-ai-reset-rewiring-to-turn-potential-into-value-in-2024


Taker
Usuario de herramientas disponibles, via APIs y servicios de suscripción
Shaper
Integrador de modelos disponibles con data propietaria
Maker
Constructor de Modelos de Lenguaje de Gran Tamaño (LLMs)
Mayor Valor



Principales Conclusiones

1. La industria avanza por delante del mundo académico.
2. Saturación de rendimiento en los benchmarks tradicionales.
3. **La IA ayuda y perjudica al medio ambiente al mismo tiempo.**
4. El mejor científico emergente del mundo... ¿la IA?
5. El número de incidentes relacionados con el uso indebido de la IA está aumentando rápidamente.
6. La demanda de habilidades profesionales relacionadas con la IA está aumentando en prácticamente todos los sectores industriales.
7. Por primera vez en la última década, la inversión privada interanual en IA disminuyó.
8. Si bien la proporción de empresas que adoptan IA se ha estancado, las empresas que la han adoptado siguen avanzando.
9. El interés de los responsables políticos en la IA está aumentando.
10. Los ciudadanos chinos se encuentran entre aquellos que tienen una opinión más positiva sobre los productos y servicios de inteligencia artificial. Los estadounidenses... no tanto.

https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf

Escalabilidad Tecnológica

Escalabilidad Tecnológica

Modelos de Lenguaje de Gran Tamaño (LLMs)

GPT-4 1.8 trillions

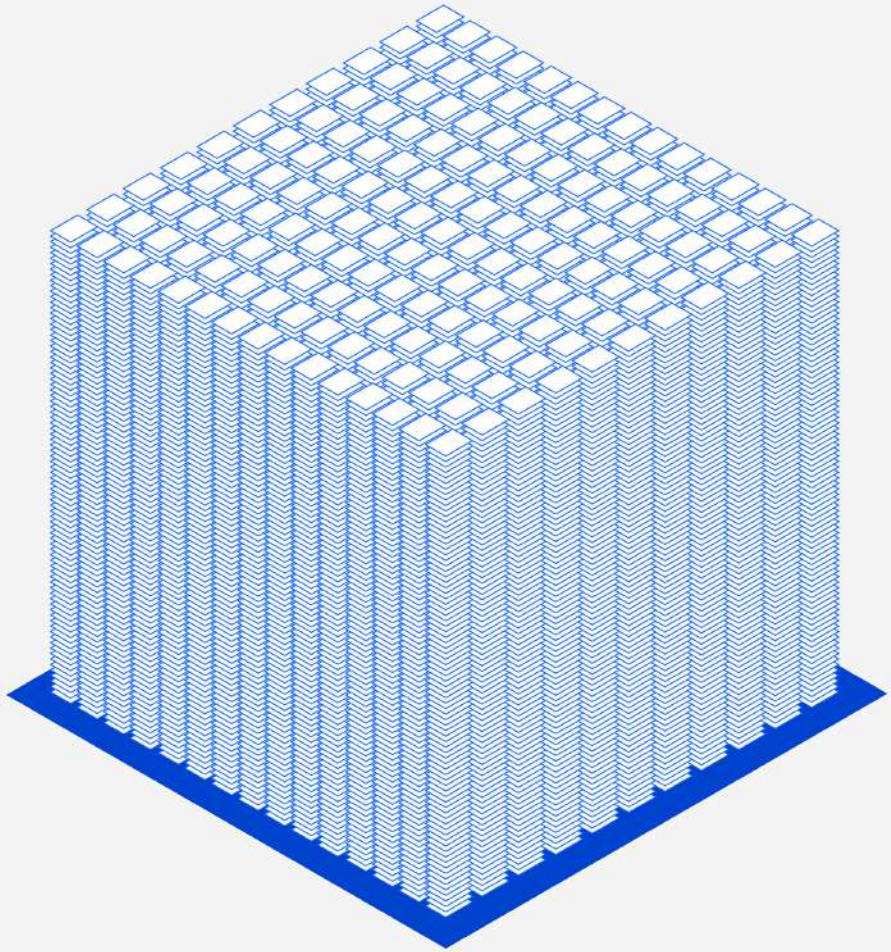
Llama 3.1 405 billions

Mixtral 8x22B 141 billions

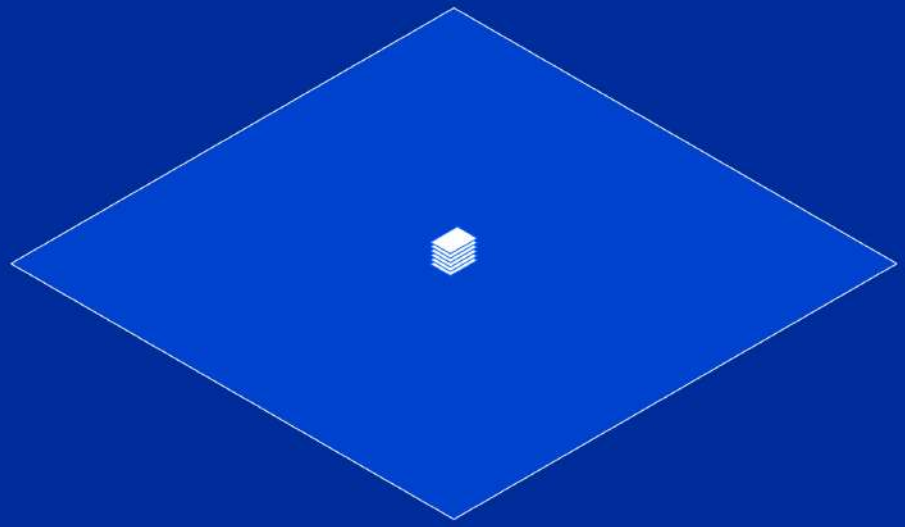
Megatron-Turing NLG 530B, costó entrenarlo alrededor de $11 millones de dólares*

*Artificial Intelligence Index Report 2023, Stanford Univeristy Human-Centered Artificial Intelligence
https://arxiv.org/pdf/2310.03715

Casi todos los datos públicos disponibles ahora están representados en modelos.

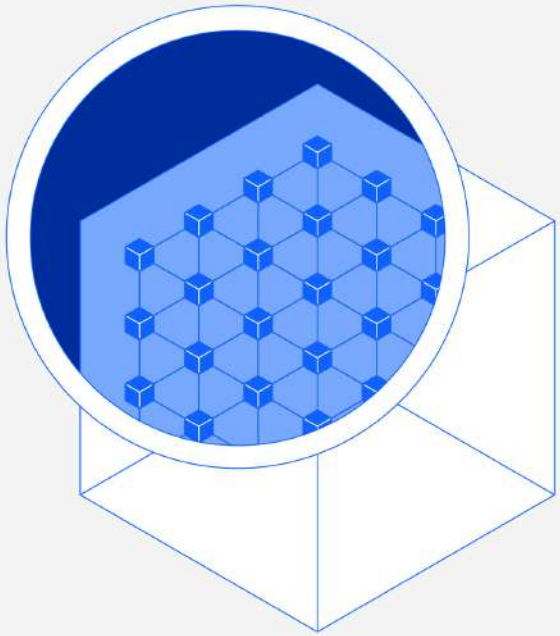


Menos del 1% de todos los datos empresariales están representados en modelos.

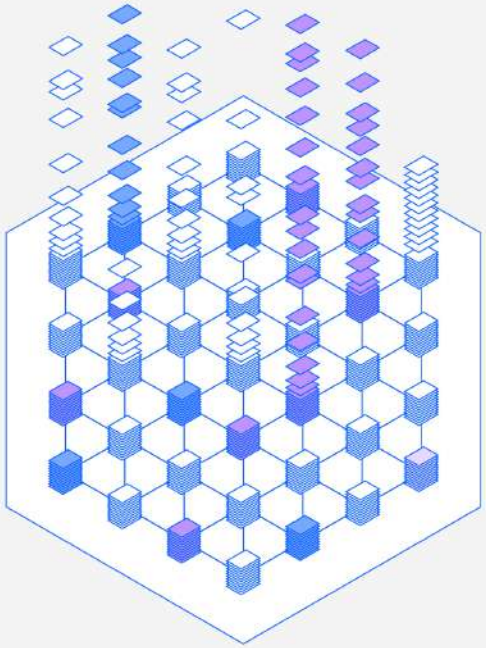


Confianza en lo que interactuamos

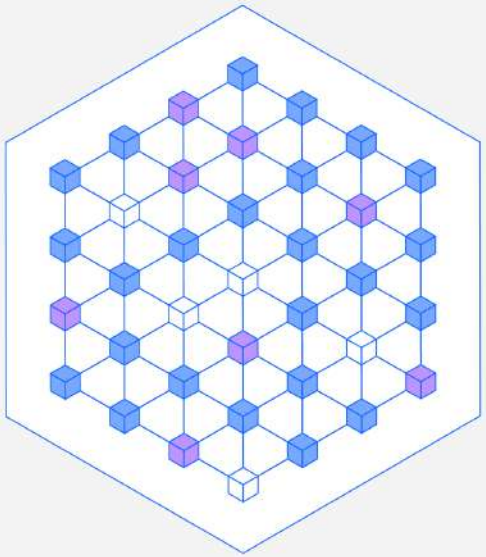
→ El modelo base debe ser transparente para que sepas cuál es su contenido.



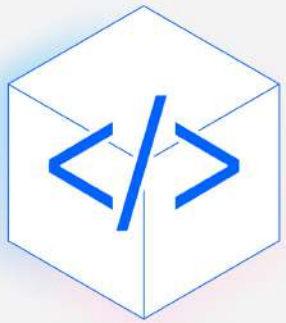
→ Porque al agregar tus datos, necesitas saber con qué los estás mezclando.



→ La combinación resultante es lo que consumirá tu empresa y tu.



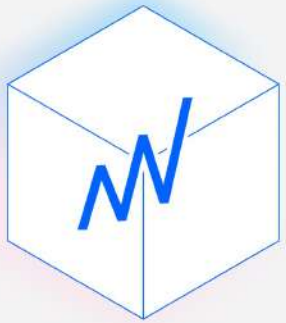
18 modelos fundacionales



IBM Granite Code models

Base
Granite-34B-Code-Base
Granite-20B-Code-Base
Granite-8B-Code-Base
Granite-3B-Code-Base

Instruction-tuned
Granite-34B-Code-Instruct
Granite-20B-Code-Instruct
Granite-8B-Code-Instruct
Granite-3B-Code-Instruct



IBM Granite Time Series models

Granite-TimeSeries-TTM
Granite-TimeSeries-PatchTSMixer
Granite-TimeSeries-PatchTST

IBM Granite Language models

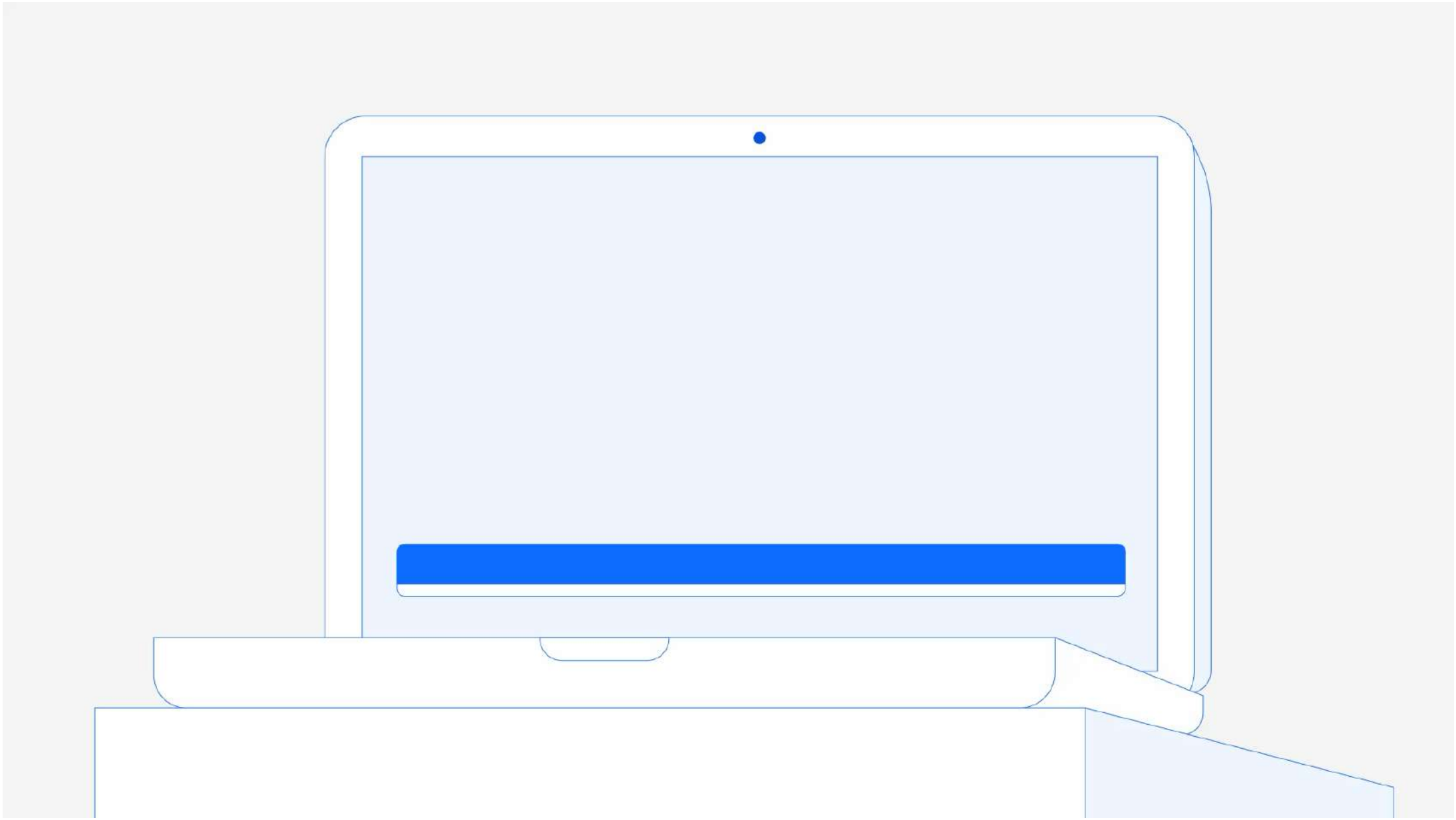
English Base
Granite-7B-Base

English Instruction-tuned
Granite-7B-Instruct

IBM Granite 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

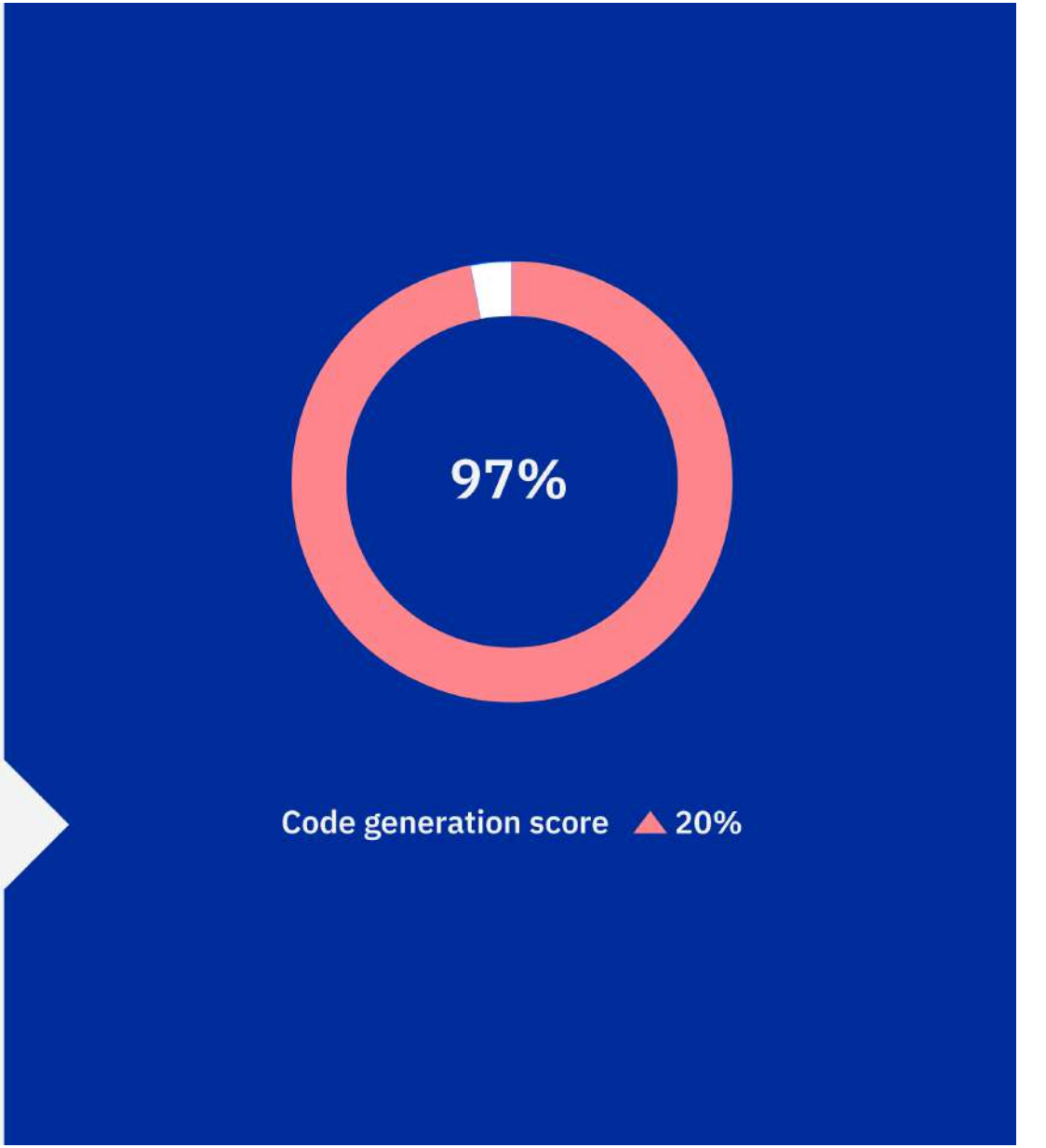


COBOL a Java con métodos existentes

→ 9 meses
→ 14 iteraciones

COBOL a Java con InstructLab

→ 1 semana
→ 1 iteración



El Futuro de la Inteligencia Artificial Empresarial

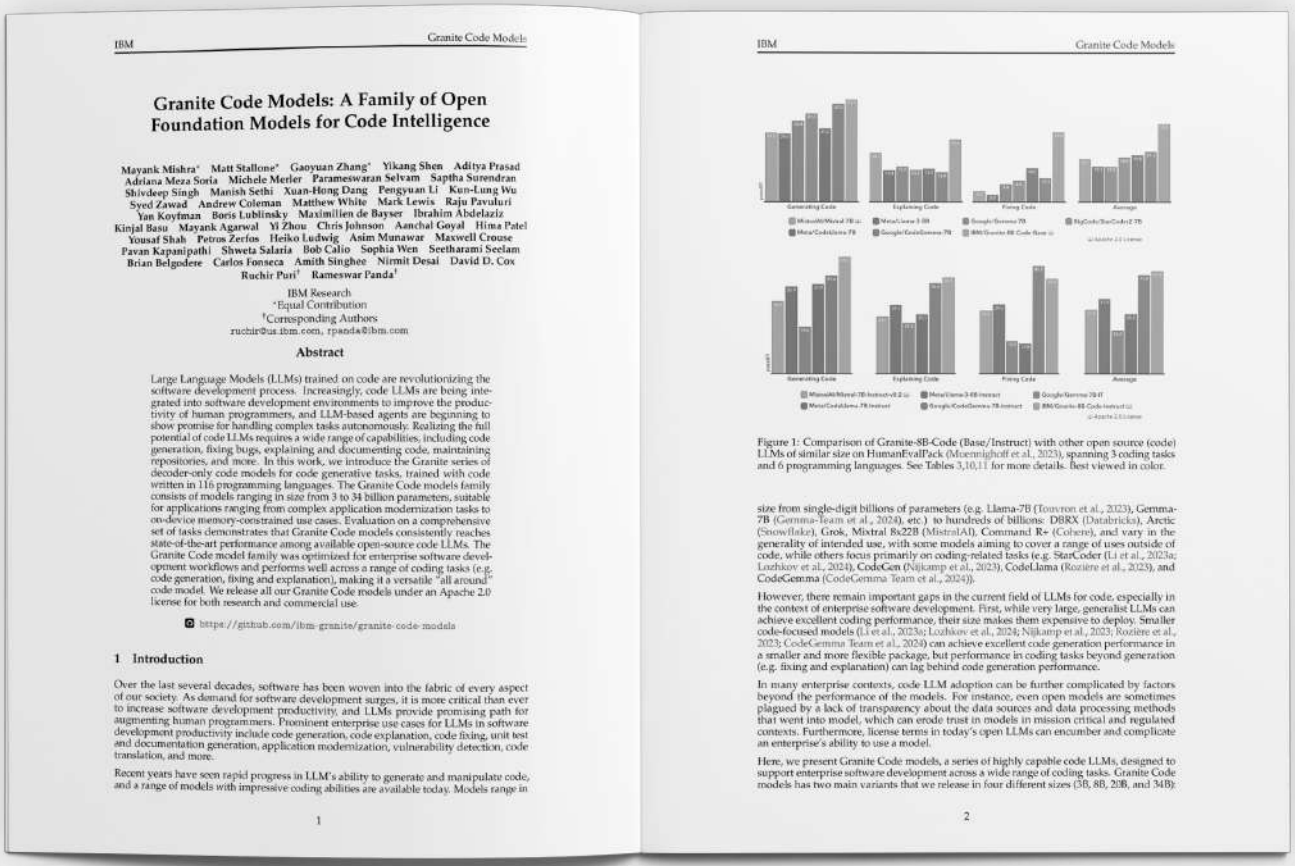
LLMs
Escalabilidad

Open
Colaboración

Quantum
Límites

Granite Code Models: A Family of Open Foundation Models for Code Intelligence

→ arXiv:2405.04324



Granite Code Models: A Family of Open Foundation Models for Code Intelligence

Mayank Mishra* Matt Stallone* Gaoyuan Zhang* Yikang Shen Aditya Prasad Adriana Meza Soria Michele Merler Parameswaran Selvam Saptha Surendran Shivdeep Singh Manish Sethi Xuan-Hong Dang Pengyuan Li Kun-Lung Wu Syed Zawad Andrew Coleman Matthew White Mark Lewis Raju Pavuluri Yan Koyfman Boris Lublinsky Maximilien de Bayser Ibrahim Abdelaziz Kinjal Basu Mayank Agarwal Yi Zhou Chris Johnson Aanchal Goyal Hima Patel Yousaf Shah Petros Zerfos Heiko Ludwig Asim Munawar Maxwell Crouse Pavan Kapanipathi Shweta Salaria Bob Calio Sophia Wen Seetharami Seelam Brian Belgodere Carlos Fonseca Amith Singhee Nirmit Desai David D. Cox Ruchir Puri† Rameswar Panda†

IBM Research
*Equal Contribution
†Corresponding Authors
ruchir@us.ibm.com, rpanda@ibm.com

Abstract

Large Language Models (LLMs) trained on code are revolutionizing the software development process. Increasingly, code LLMs are being integrated into software development environments to improve the productivity of human programmers, and LLM-based agents are beginning to show promise for handling complex tasks autonomously. Realizing the full potential of code LLMs requires a wide range of capabilities, including code generation, fixing bugs, explaining and documenting code, maintaining repositories, and more. In this work, we introduce the Granite series of decoder-only code models for code generative tasks, trained with code written in 116 programming languages. The Granite Code models family consists of models ranging in size from 3 to 34 billion parameters, suitable for applications ranging from complex application modernization tasks to on-device memory-constrained use cases. Evaluation on a comprehensive set of tasks demonstrates that Granite Code models consistently reaches state-of-the-art performance among available open-source code LLMs. The Granite Code model family was optimized for enterprise software development workflows and performs well across a range of coding tasks (e.g. code generation, fixing and explanation), making it a versatile "all around" code model. We release all our Granite Code models under an Apache 2.0 license for both research and commercial use.

https://github.com/ibm-granite/granite-code-models

1 Introduction

Over the last several decades, software has been woven into the fabric of every aspect of our society. As demand for software development surges, it is more critical than ever to increase software development productivity, and LLMs provide promising path for augmenting human programmers. Prominent enterprise use cases for LLMs in software development productivity include code generation, code explanation, code fixing, unit test and documentation generation, application modernization, vulnerability detection, code translation, and more.

Recent years have seen rapid progress in LLM's ability to generate and manipulate code, and a range of models with impressive coding abilities are available today. Models range in

Figure 1: Comparison of Granite-8B-Code (Base/Instruct) with other open source (code) LLMs of similar size on HumanEvalPack (Muennighoff et al., 2023), spanning 3 coding tasks and 6 programming languages. See Tables 3,10,11 for more details. Best viewed in color.

size from single-digit billions of parameters (e.g. Llama-7B (Touvron et al., 2023), Gemma-7B (Gemma-Team et al., 2024), etc.) to hundreds of billions: DBRX (Databricks), Arctic (Snowflake), Grok, Mixtral 8x22B (MistralAI), Command R+ (Cohere), and vary in the generality of intended use, with some models aiming to cover a range of uses outside of code, while others focus primarily on coding-related tasks (e.g. StarCoder (Li et al., 2023a; Lozhkov et al., 2024), CodeGen (Nijkamp et al., 2023), CodeLlama (Rozière et al., 2023), and CodeGemma (CodeGemma Team et al., 2024)).

However, there remain important gaps in the current field of LLMs for code, especially in the context of enterprise software development. First, while very large, generalist LLMs can achieve excellent coding performance, their size makes them expensive to deploy. Smaller code-focused models (Li et al., 2023a; Lozhkov et al., 2024; Nijkamp et al., 2023; Rozière et al., 2023; CodeGemma Team et al., 2024) can achieve excellent code generation performance in a smaller and more flexible package, but performance in coding tasks beyond generation (e.g. fixing and explanation) can lag behind code generation performance.

In many enterprise contexts, code LLM adoption can be further complicated by factors beyond the performance of the models. For instance, even open models are sometimes plagued by a lack of transparency about the data sources and data processing methods that went into model, which can erode trust in models in mission critical and regulated contexts. Furthermore, license terms in today's open LLMs can encumber and complicate an enterprise's ability to use a model.

Here, we present Granite Code models, a series of highly capable code LLMs, designed to support enterprise software development across a wide range of coding tasks. Granite Code models has two main variants that we release in four different sizes (3B, 8B, 20B, and 34B):

Construcción de un Modelo Fundacional de IA para el Tiempo & Clima

NASA & IBM

Predicción Clima
6 días / 2024

5 días / 2014

2023

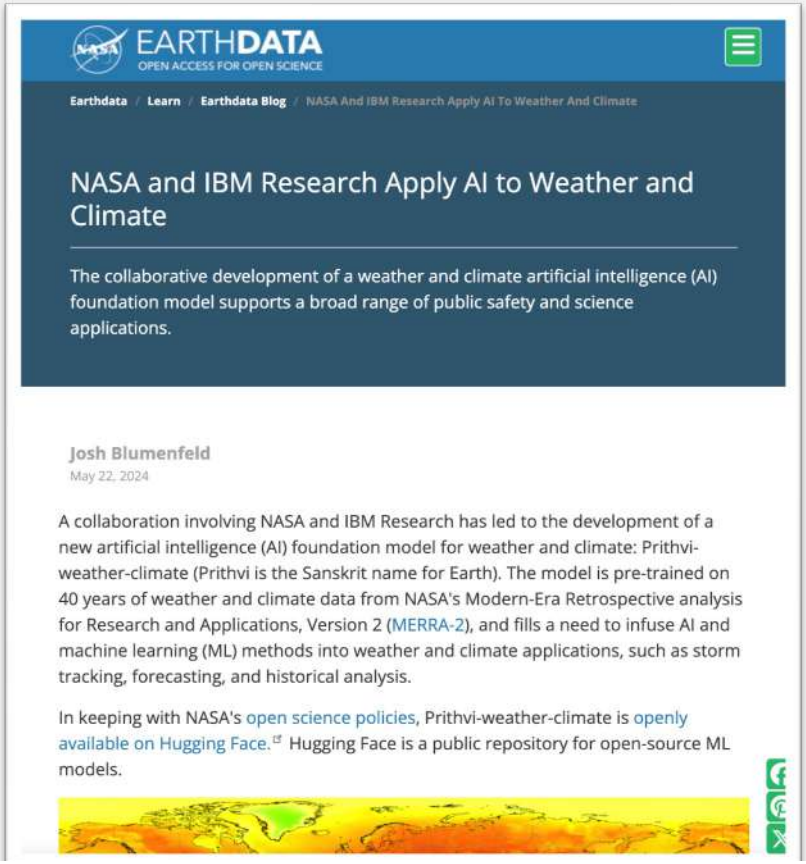
Modelo Geospacial

Satélite **Harmonized Landsat Sentinel-2** (HLS) de NASA
- Imágenes de EEUU
- 1 año de información

2024

Modelo Prithvi WxC

Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)
- Mundo entero
- 40 años de información
- Entrenamiento continuo


EARTHDATA
OPEN ACCESS FOR OPEN SCIENCE

Earthdata / Learn / Earthdata Blog / NASA And IBM Research Apply AI To Weather And Climate

NASA and IBM Research Apply AI to Weather and Climate

The collaborative development of a weather and climate artificial intelligence (AI) foundation model supports a broad range of public safety and science applications.

Josh Blumenfeld
May 22, 2024

A collaboration involving NASA and IBM Research has led to the development of a new artificial intelligence (AI) foundation model for weather and climate: Prithvi-weather-climate (Prithvi is the Sanskrit name for Earth). The model is pre-trained on 40 years of weather and climate data from NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2), and fills a need to infuse AI and machine learning (ML) methods into weather and climate applications, such as storm tracking, forecasting, and historical analysis.

In keeping with NASA's open science policies, Prithvi-weather-climate is openly available on Hugging Face. Hugging Face is a public repository for open-source ML models.

Gracias a la Inteligencia Artificial se descubren más de 300 geoglifos en Nazca, Perú

Universidad de Yamagata & IBM Research

1940s

112 figuras
39 figuras de línea
73 figuras de alto relieve

10 años

2004 - 2020

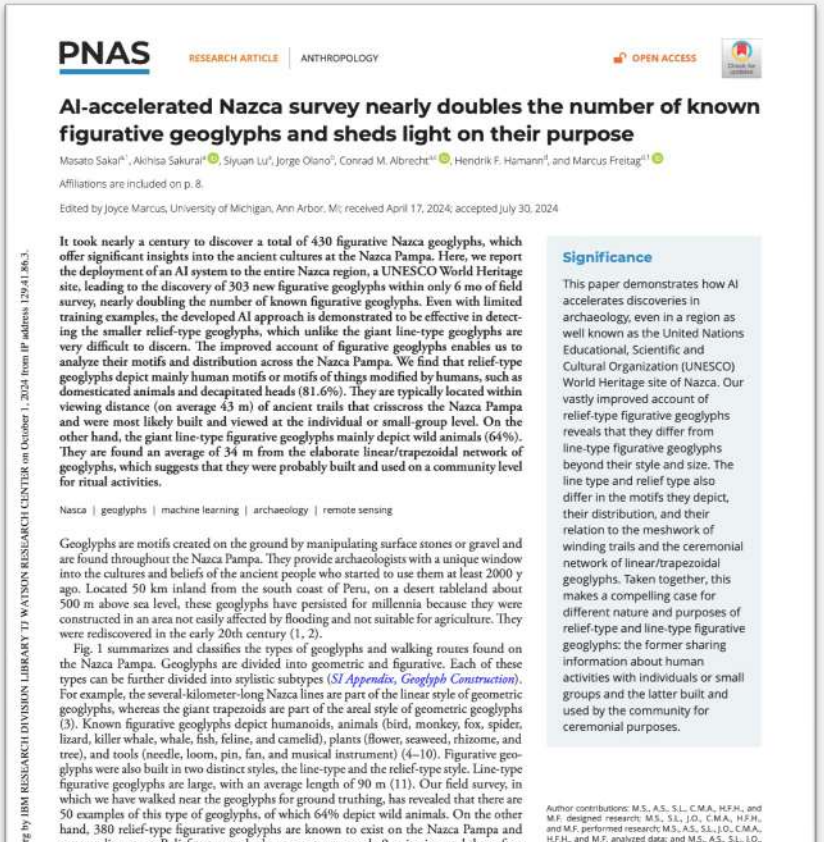
318 figuras
11 figuras de línea
307 figuras de alto relieve

16 años

2024

303 figuras
303 figuras de alto relieve

6 meses



PNAS RESEARCH ARTICLE | ANTHROPOLOGY OPEN ACCESS

AI-accelerated Nazca survey nearly doubles the number of known figurative geoglyphs and sheds light on their purpose

Masato Sakai[a,1], Akihisa Sakurai[a], Siyuan Lu[a], Jorge Olano[b], Conrad M. Albrecht[a,c], Hendrik F. Hamann[d], and Marcus Freitag[d,1]

Affiliations are included on p. 8.

Edited by Joyce Marcus, University of Michigan, Ann Arbor, MI; received April 17, 2024; accepted July 30, 2024

It took nearly a century to discover a total of 430 figurative Nazca geoglyphs, which offer significant insights into the ancient cultures at the Nazca Pampa. Here, we report the deployment of an AI system to the entire Nazca region, a UNESCO World Heritage site, leading to the discovery of 303 new figurative geoglyphs within only 6 mo of field survey, nearly doubling the number of known figurative geoglyphs. Even with limited training examples, the developed AI approach is demonstrated to be effective in detecting the smaller relief-type geoglyphs, which unlike the giant line-type geoglyphs are very difficult to discern. The improved account of figurative geoglyphs enables us to analyze their motifs and distribution across the Nazca Pampa. We find that relief-type geoglyphs depict mainly human motifs or motifs of things modified by humans, such as domesticated animals and decapitated heads (81.6%). They are typically located within viewing distance (on average 43 m) of ancient trails that crisscross the Nazca Pampa and were most likely built and viewed at the individual or small-group level. On the other hand, the giant line-type figurative geoglyphs mainly depict wild animals (64%). They are found an average of 34 m from the elaborate linear/trapezoidal network of geoglyphs, which suggests that they were probably built and used on a community level for ritual activities.

Nasca | geoglyphs | machine learning | archaeology | remote sensing

Significance

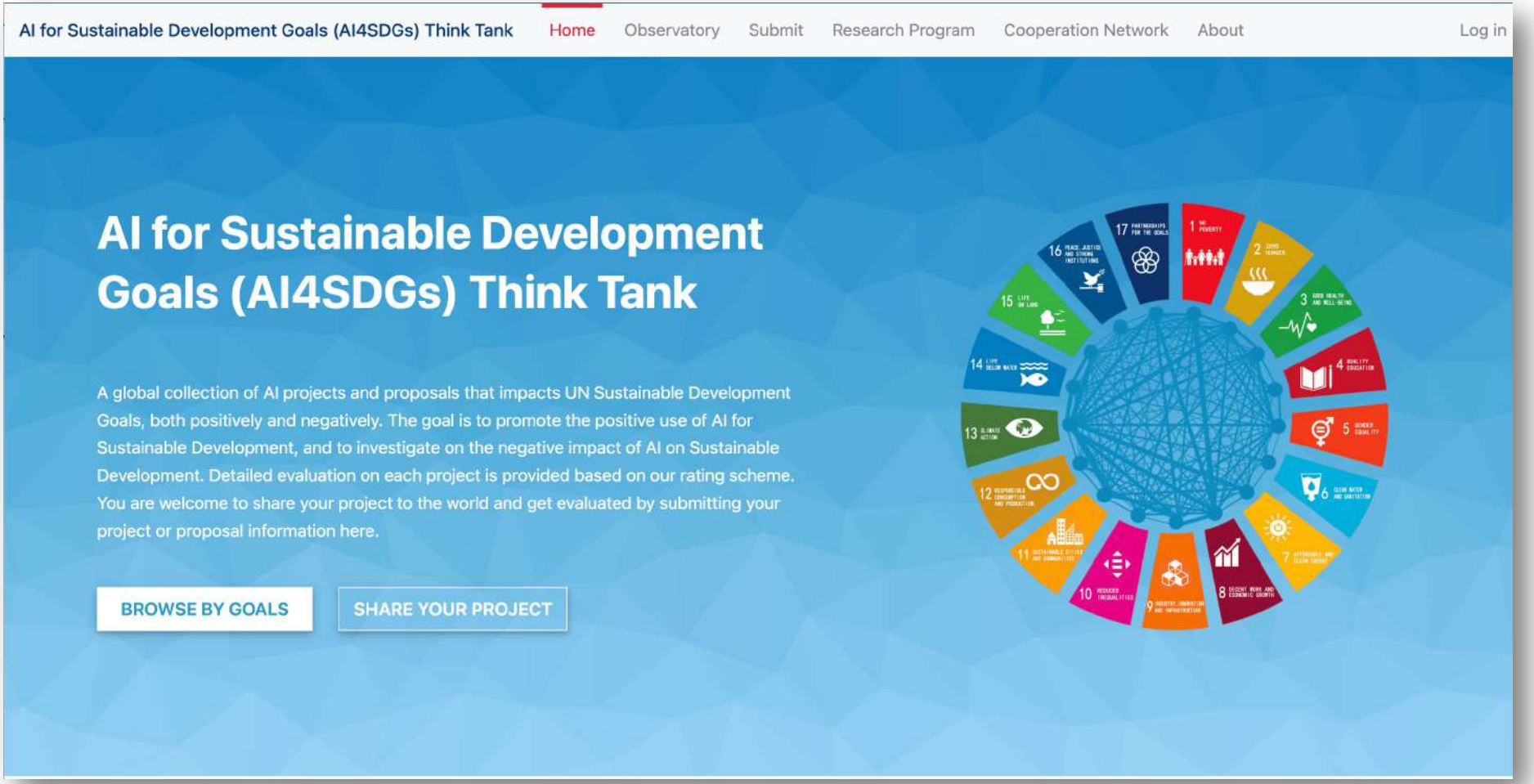
This paper demonstrates how AI accelerates discoveries in archaeology, even in a region as well known as the United Nations Educational, Scientific and Cultural Organization (UNESCO) World Heritage site of Nazca. Our vastly improved account of relief-type figurative geoglyphs reveals that they differ from line-type figurative geoglyphs beyond their style and size. The line type and relief type also differ in the motifs they depict, their distribution, and their relation to the meshwork of winding trails and the ceremonial network of linear/trapezoidal geoglyphs. Taken together, this makes a compelling case for different nature and purposes of relief-type and line-type figurative geoglyphs: the former sharing information about human activities with individuals or small groups and the latter built and used by the community for ceremonial purposes.

Geoglyphs are motifs created on the ground by manipulating surface stones or gravel and are found throughout the Nazca Pampa. They provide archaeologists with a unique window into the cultures and beliefs of the ancient people who started to use them at least 2000 y ago. Located 50 km inland from the south coast of Peru, on a desert tableland about 500 m above sea level, these geoglyphs have persisted for millennia because they were constructed in an area not easily affected by flooding and not suitable for agriculture. They were rediscovered in the early 20th century (1, 2).

Fig. 1 summarizes and classifies the types of geoglyphs and walking routes found on the Nazca Pampa. Geoglyphs are divided into geometric and figurative. Each of these types can be further divided into stylistic subtypes (*SI Appendix*, *Geoglyph Construction*). For example, the several-kilometer-long Nazca lines are part of the linear style of geometric geoglyphs, whereas the giant trapezoids are part of the areal style of geometric geoglyphs (3). Known figurative geoglyphs depict humanoids, animals (bird, monkey, fox, spider, lizard, killer whale, whale, fish, feline, and camelid), plants (flower, seaweed, rhizome, and tree), and tools (needle, loom, pin, fan, and musical instrument) (4–10). Figurative geoglyphs were also built in two distinct styles, the line-type and the relief-type style. Line-type figurative geoglyphs are large, with an average length of 90 m (11). Our field survey, in which we have walked near the geoglyphs for ground truthing, has revealed that there are 50 examples of this type of geoglyphs, of which 64% depict wild animals. On the other hand, 380 relief-type figurative geoglyphs are known to exist on the Nazca Pampa and

Author contributions: M.S., A.S., S.L., C.M.A., H.F.H., and M.F. designed research; M.S., S.L., J.O., C.M.A., H.F.H., and M.F. performed research; M.S., A.S., S.L., J.O., C.M.A., H.F.H., and M.F. analyzed data; and M.S., A.S., S.L., J.O.,

The AI Alliance



https://ai-for-sdgs.academy/

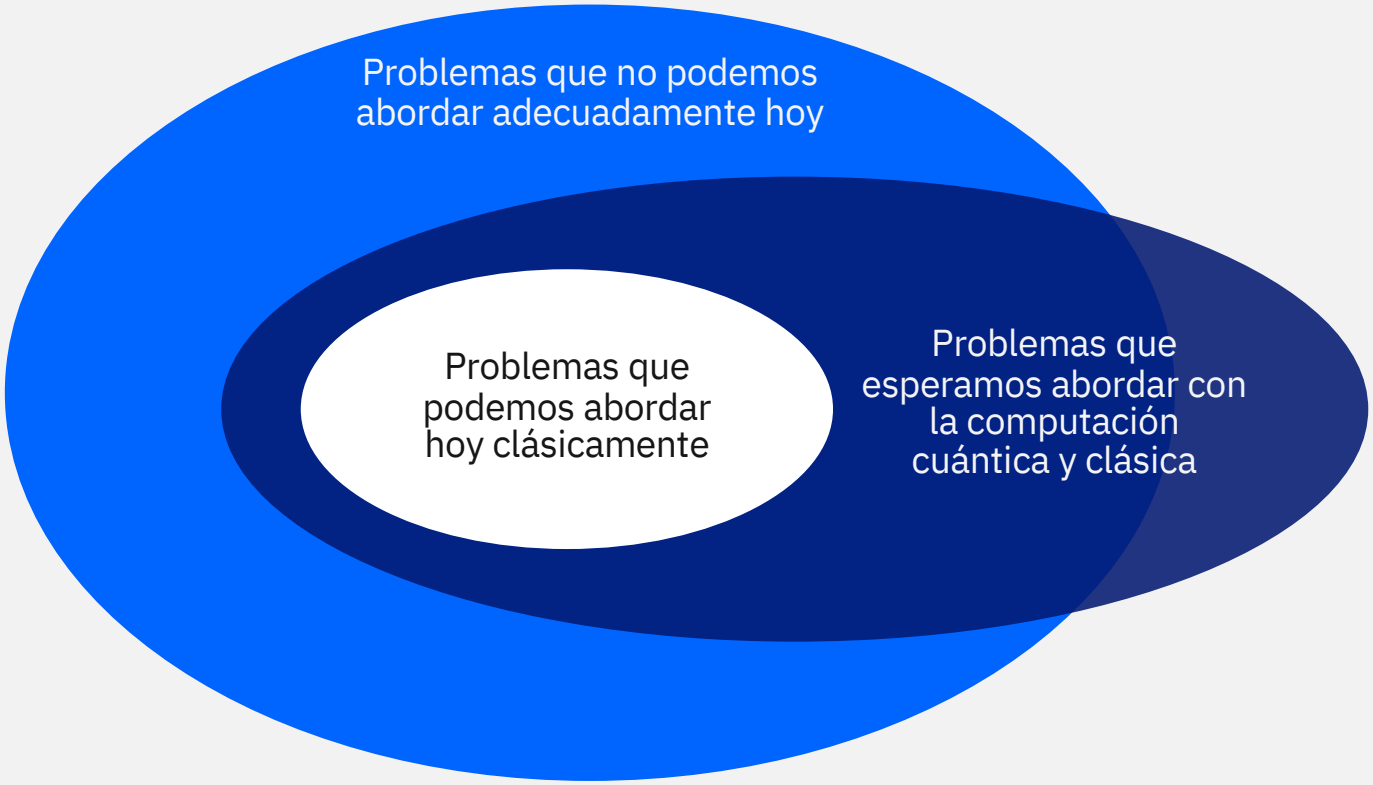
El Futuro de la Inteligencia Artificial Empresarial

LLMs
Escalabilidad

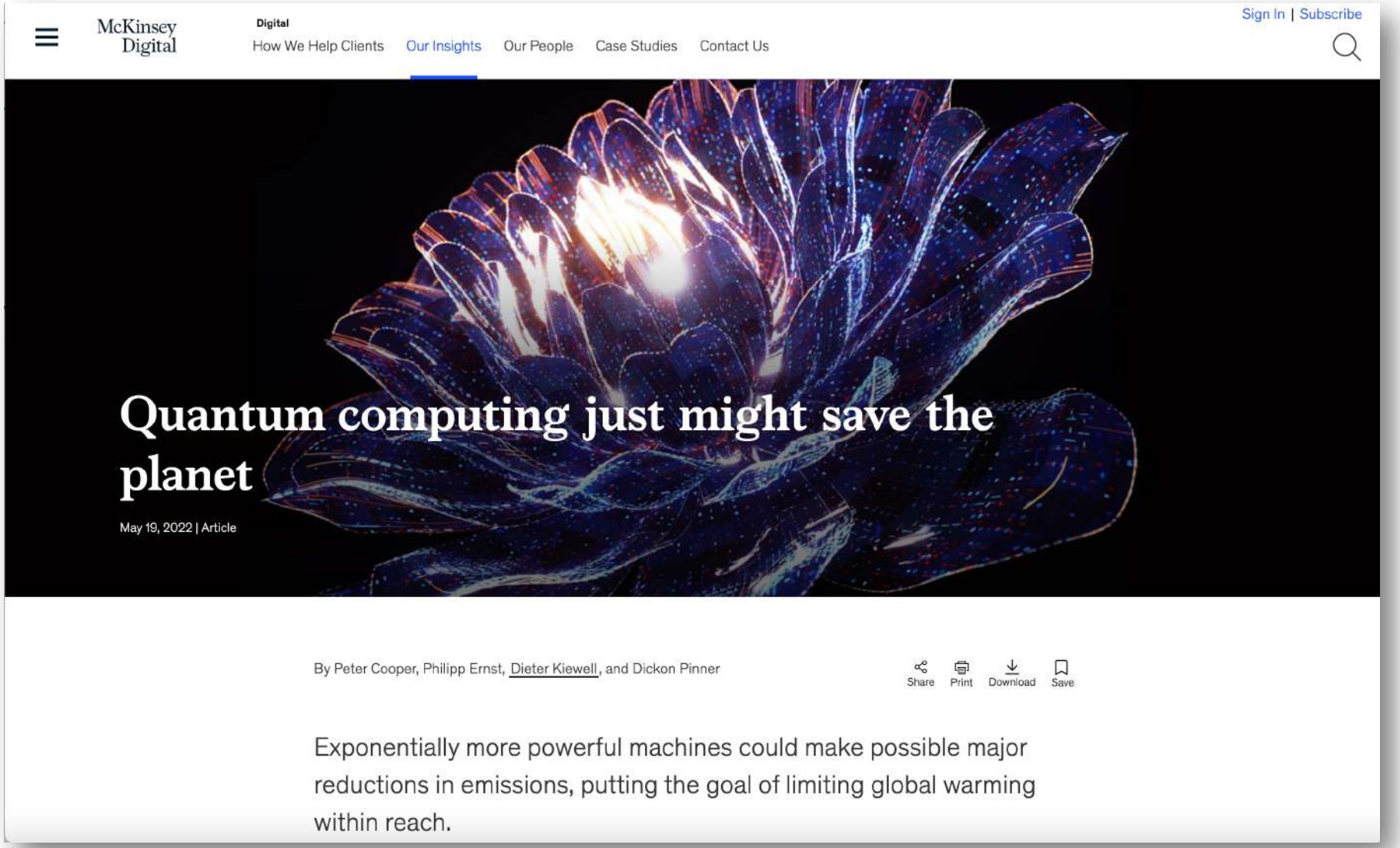
Open
Colaboración

Quantum
Límites

¿Por qué computación cuántica?



A pesar de lo sofisticada que se ha vuelto la informática digital "clásica", hay muchos problemas científicos y comerciales para los que apenas tenemos respuesta.



https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/quantum-computing-just-might-save-the-planet

Las aplicaciones cuánticas abarcan tres áreas generales

Sistemas Cuánticos

- Química cuántica
- Ciencia de Materiales
- Física de alta energía

Inteligencia Artificial

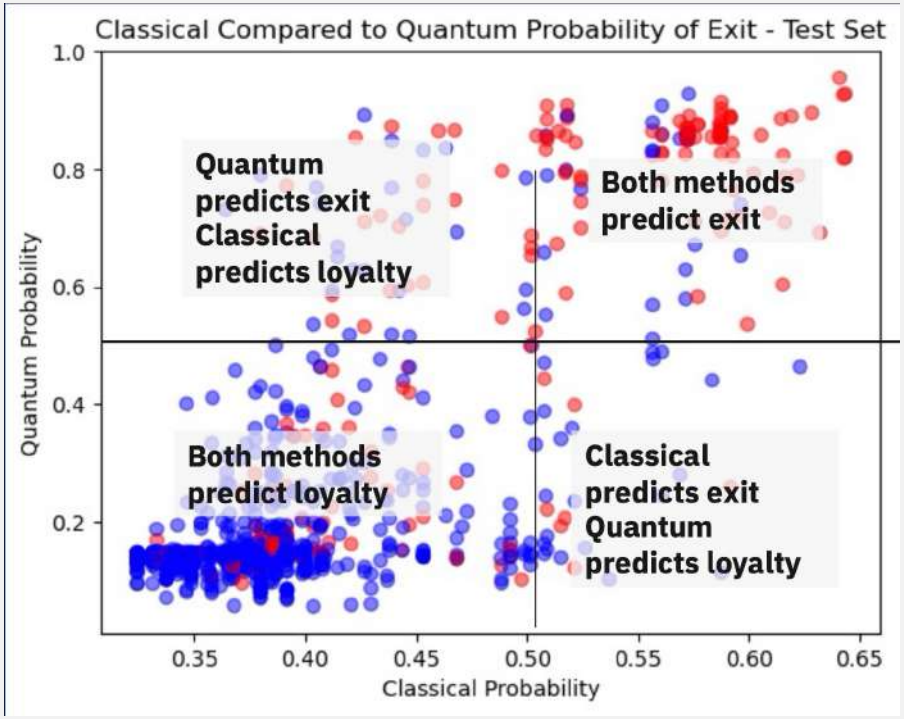
- Mejor entrenamiento de modelos
- Reconocimiento de patrones
- Detección de fraude

Optimización

- Optimización de cartera
- Análisis de riesgo
- Préstamos y puntuación de crédito
- Aplicaciones tipo Monte Carlo

Algoritmos Híbridos Cuánticos & Clásicos

Comportamiento de Clientes

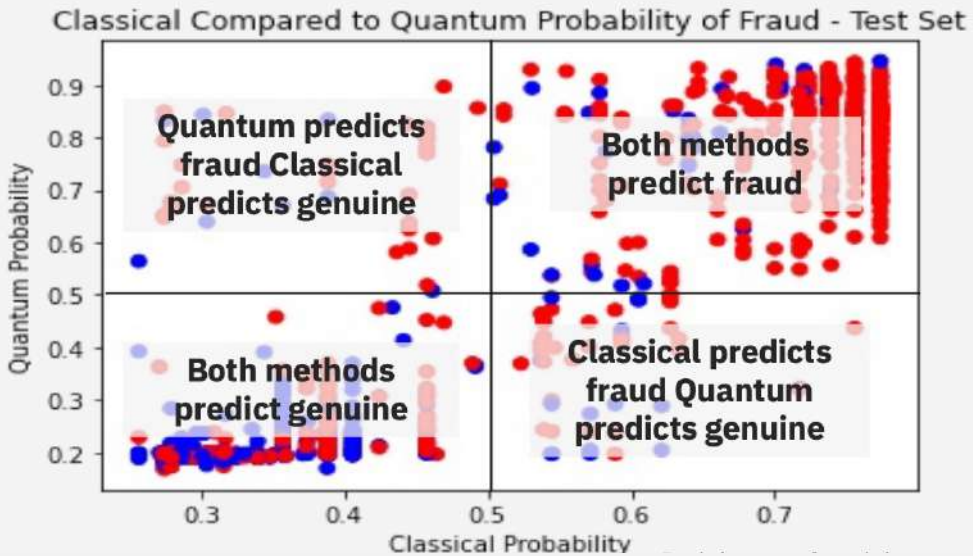


Puntos Rojos: Clientes que se van

Puntos Azules: Clientes que son leales

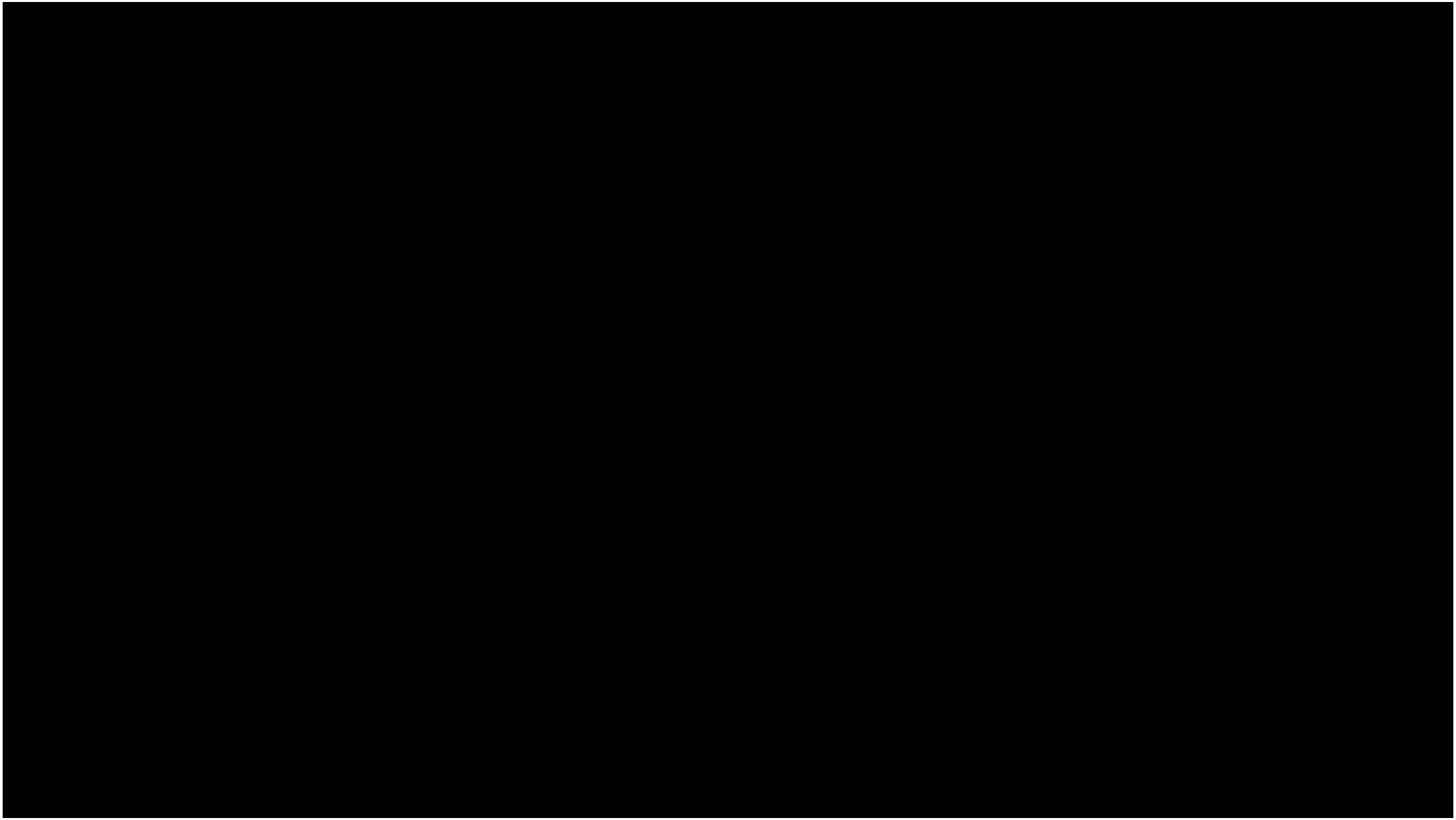
Algoritmo de Detección de Fraude en Pagos

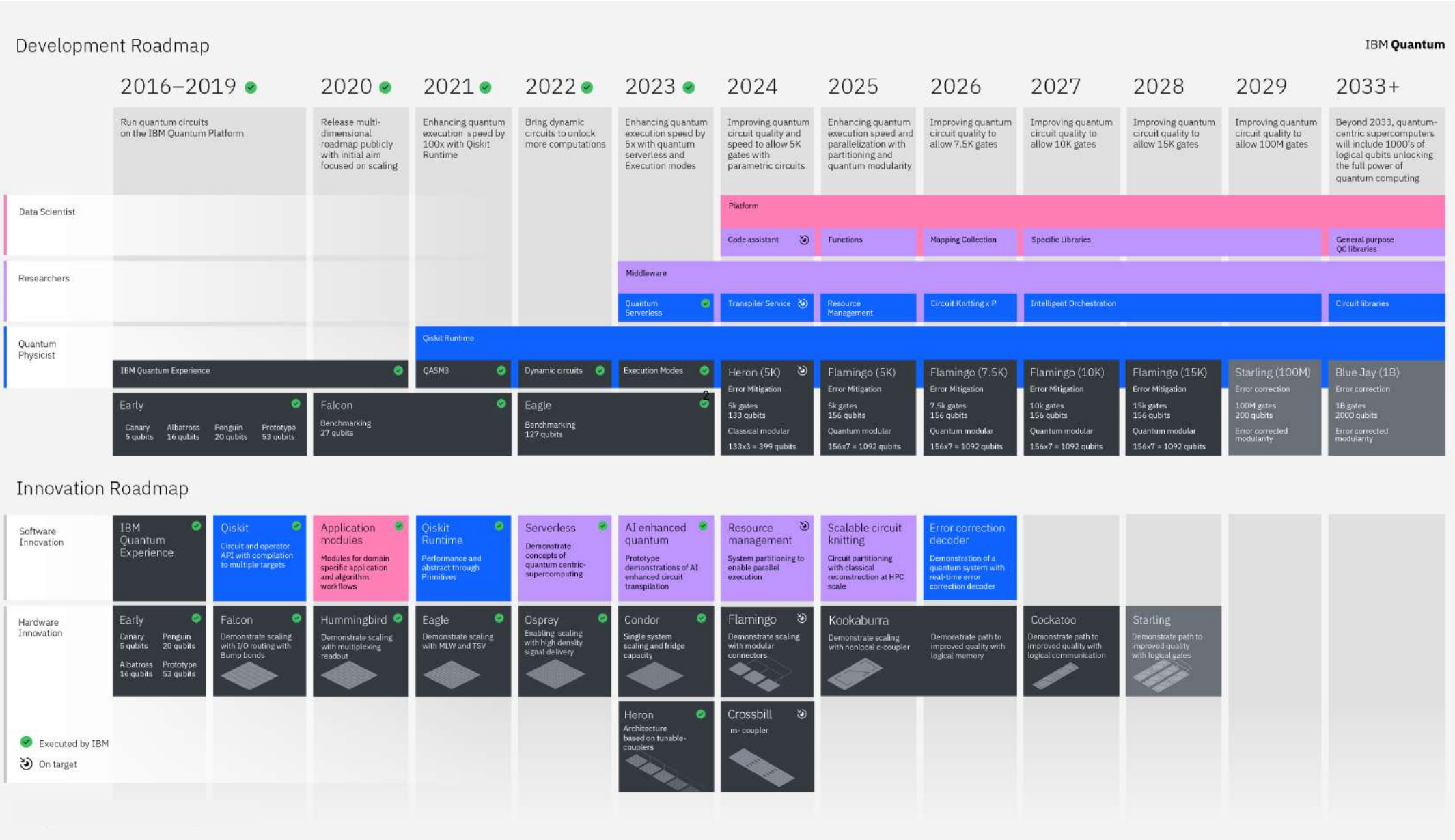
Complementation of Classical + Quantum:



Puntos Rojos: Transacciones fraudulentas

Puntos Azules: Transacciones válidas




Development Roadmap
IBM Quantum
2016–2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2033+
Run quantum circuits on the IBM Quantum Platform
Release multi-dimensional roadmap publicly with initial aim focused on scaling
Enhancing quantum execution speed by 100x with Qiskit Runtime
Bring dynamic circuits to unlock more computations
Enhancing quantum execution speed by 5x with quantum serverless and Execution modes
Improving quantum circuit quality and speed to allow 5K gates with parametric circuits
Enhancing quantum execution speed and parallelization with partitioning and quantum modularity
Improving quantum circuit quality to allow 7.5K gates
Improving quantum circuit quality to allow 10K gates
Improving quantum circuit quality to allow 15K gates
Improving quantum circuit quality to allow 100M gates
Beyond 2033, quantum-centric supercomputers will include 1000's of logical qubits unlocking the full power of quantum computing
Data Scientist
Platform
Code assistant
Functions
Mapping Collection
Specific Libraries
General purpose QC libraries
Researchers
Middleware
Quantum Serverless
Transpiler Service
Resource Management
Circuit Knitting x P
Intelligent Orchestration
Circuit libraries
Quantum Physicist
Qiskit Runtime
IBM Quantum Experience
QASM3
Dynamic circuits
Execution Modes
Heron (5K)
Error Mitigation
5k gates
133 qubits
Classical modular
133x3 = 399 qubits
Flamingo (5K)
Error Mitigation
5k gates
156 qubits
Quantum modular
156x7 = 1092 qubits
Flamingo (7.5K)
Error Mitigation
7.5k gates
156 qubits
Quantum modular
156x7 = 1092 qubits
Flamingo (10K)
Error Mitigation
10k gates
156 qubits
Quantum modular
156x7 = 1092 qubits
Flamingo (15K)
Error Mitigation
15k gates
156 qubits
Quantum modular
156x7 = 1092 qubits
Starling (100M)
Error correction
100M gates
200 qubits
Error corrected modularity
Blue Jay (1B)
Error correction
1B gates
2000 qubits
Error corrected modularity
Early
Canary 5 qubits
Albatross 16 qubits
Penguin 20 qubits
Prototype 53 qubits
Falcon
Benchmarking
27 qubits
Eagle
Benchmarking
127 qubits
Innovation Roadmap
Software Innovation
IBM Quantum Experience
Qiskit
Circuit and operator API with compilation to multiple targets
Application modules
Modules for domain specific application and algorithm workflows
Qiskit Runtime
Performance and abstract through Primitives
Serverless
Demonstrate concepts of quantum centric-supercomputing
AI enhanced quantum
Prototype demonstrations of AI enhanced circuit transpilation
Resource management
System partitioning to enable parallel execution
Scalable circuit knitting
Circuit partitioning with classical reconstruction at HPC scale
Error correction decoder
Demonstration of a quantum system with real-time error correction decoder
Hardware Innovation
Early
Canary 5 qubits
Penguin 20 qubits
Albatross 16 qubits
Prototype 53 qubits
Falcon
Demonstrate scaling with I/O routing with Bump bonds
Hummingbird
Demonstrate scaling with multiplexing readout
Eagle
Demonstrate scaling with MLW and TSV
Osprey
Enabling scaling with high density signal delivery
Condor
Single system scaling and fridge capacity
Flamingo
Demonstrate scaling with modular connectors
Kookaburra
Demonstrate scaling with nonlocal c-coupler
Demonstrate path to improved quality with logical memory
Cockatoo
Demonstrate path to improved quality with logical communication
Starling
Demonstrate path to improved quality with logical gates
Heron
Architecture based on tunable-couplers
Crossbill
m- coupler
Executed by IBM
On target

El Futuro de la Inteligencia Artificial Empresarial

LLMs
Escalabilidad

Open
Colaboración

Quantum
Límites

La Inteligencia Artificial para un Futuro Sostenible

Narciso Lema
CTO & Technical Community Leader IBM
Colombia, Perú, Ecuador, Venezuela & LCR
https://www.linkedin.com/in/narcisolema/