IA Generativa, el equilibrio entre subestimar y sobrestimar

Ing. Eduardo Mangarelli

mangarelli@ort.edu.uy

El fenómeno actual de los sistemas de Inteligencia Artificial Generativa

Nos puede llevar a hablar de…

Múltiples dimensiones de análisis

Tecnológica
- Modelos/LLMs
- Infraestructura
- Datos
- Cómputo
- Empresas

Económica
- Productividad
- Mercados públicos de capital
- Inversión privada / VC
- Energía

Laboral
- Impacto en el empleo
- Automatización
- Nuevas habilidades

Organizacional
- Rediseño de procesos
- Nuevas formas de trabajo
- Gobernanza y gestión del cambio

Educativa
- Rediseño curricular
- Impacto para docentes y alumnos

Social y cultural
- Impacto social
- Acceso y desigualdad
- Sesgos

Política e institucional
- Regulación y derecho
- Políticas públicas

Geopolítica y Ambiental
- Competencia entre países
- Soberanía tecnológica
- Recursos e impacto ecológico

Alcance & Velocidad

IA Generativa

Inteligencia Artificial

Sistemas informáticos que son capaces de *aprender* de un conjunto de *datos*, para con dicho aprendizaje:

- resolver problemas altamente complejos
- realizar predicciones
- tomar acciones automáticas
- realizar tareas identificadas como capacidades humanas

Inteligencia Artificial

~20 años de desarrollo de sistemas de uso real

- Visión por computación
- Motores de recomendación
- Optimización de rutas
- Predicciones
- Optimización de precios
- Detección de anomalías
- Segmentación
- Y más…

2017: Attention Is All You Need

Attention Is All You Need

Ashish Vaswani[*]
Google Brain
avaswani@google.com

Noam Shazeer[*]
Google Brain
noam@google.com

Niki Parmar[*]
Google Research
nikip@google.com

Jakob Uszkoreit[*]
Google Research
usz@google.com

Llion Jones[*]
Google Research
llion@google.com

Aidan N. Gomez[*] [†]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*] [‡]
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Attention Is All You Need: 2017
-> IA Generativa

Es una subdisciplina dentro de los sistemas de Inteligencia Artificial

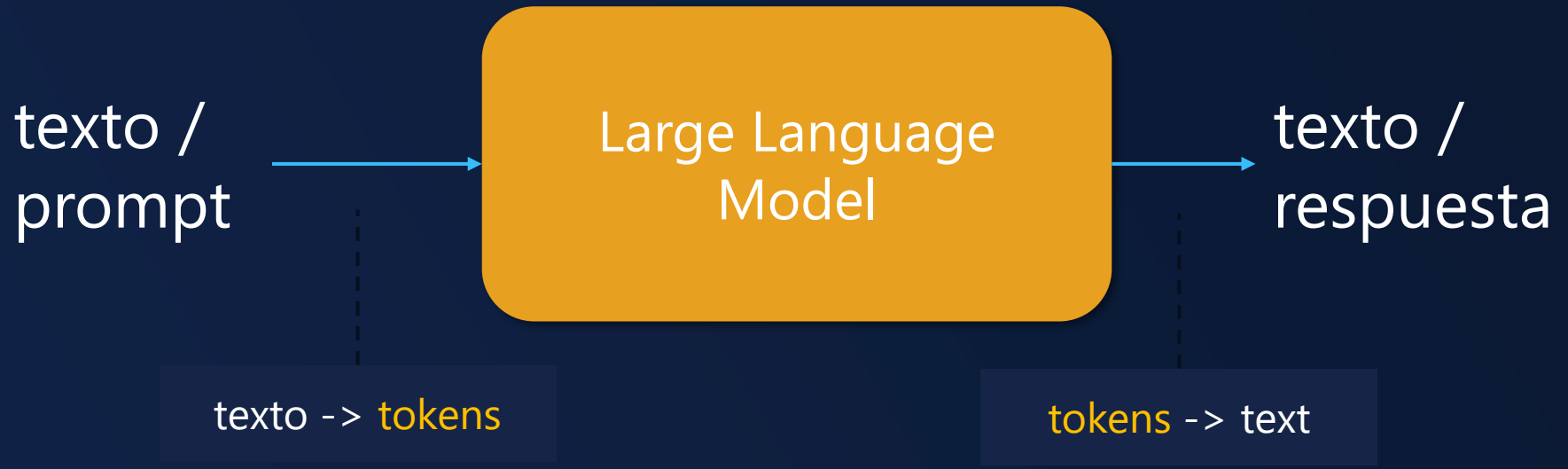
Motor detrás del un chat: Large Language Model



quiero opciones de cómo expresar que GenAI es una tecnología transversal y multipropósito

GenAI es una tecnología de propósito general, aplicable de maneras distintas según las necesidades, contextos y objetivos de cada persona u organización.

Motor detrás del chat: Large Language Model



texto: En el contexto de AI, la sigla GPT significa:
->
tokens: [1568, 650, 50124, 334, 20837, 11, 557, 6377, 2425, 174803, 33137, 25]

tokens: [5926, 1799, 4659, 118806, 113133, 13]
->
texto: Generative Pre-trained Transformer.

https://platform.openai.com/tokenizer

IA Generativa

Entrenamiento



Datos

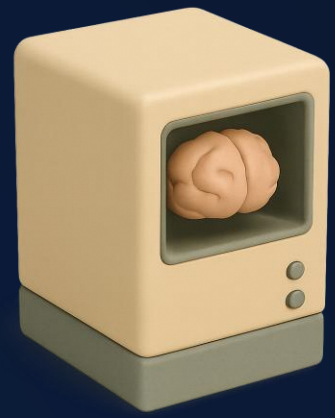


Entrenamiento
"Attention is all you need"
[GPUs]



Fine-Tuning
[GPUs]

Foundational Model



Large Language Model
(Gemini, GPT, Claude, Grok, DeepSeek...)

GPU ~= NVIDIA

IA Generativa



Datos



Entrenamiento
"Attention is all you need"
[GPUs]

Foundational Model

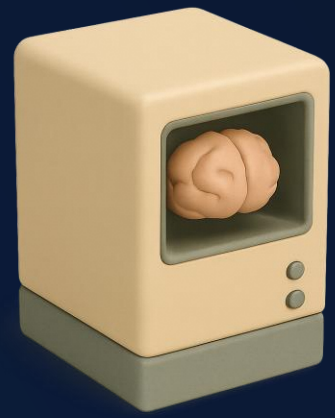


Fine-Tuning
[GPUs]



Inferencia
[GPUs]





Large Language Model
(Gemini, GPT, Claude, Grok, DeepSeek...)

GPU ~= NVIDIA

IA Generativa



Datos



Entrenamiento
"Attention is all you need"
[GPUs]



Fine-Tuning
[GPUs]

Foundational Model



ChatGPT
Gemini
Claude
Grok

Inferencia
[GPUs]





Large Language Model

Inferencia

GPU ~= NVIDIA

IA Generativa



Inferencia
[GPUs]

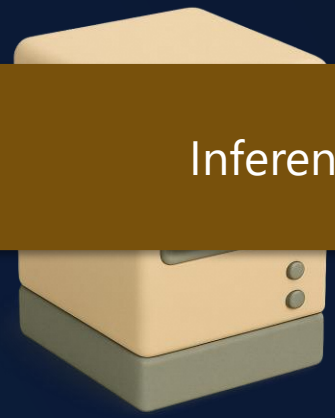












Entrenamiento

Inferencia

Datos

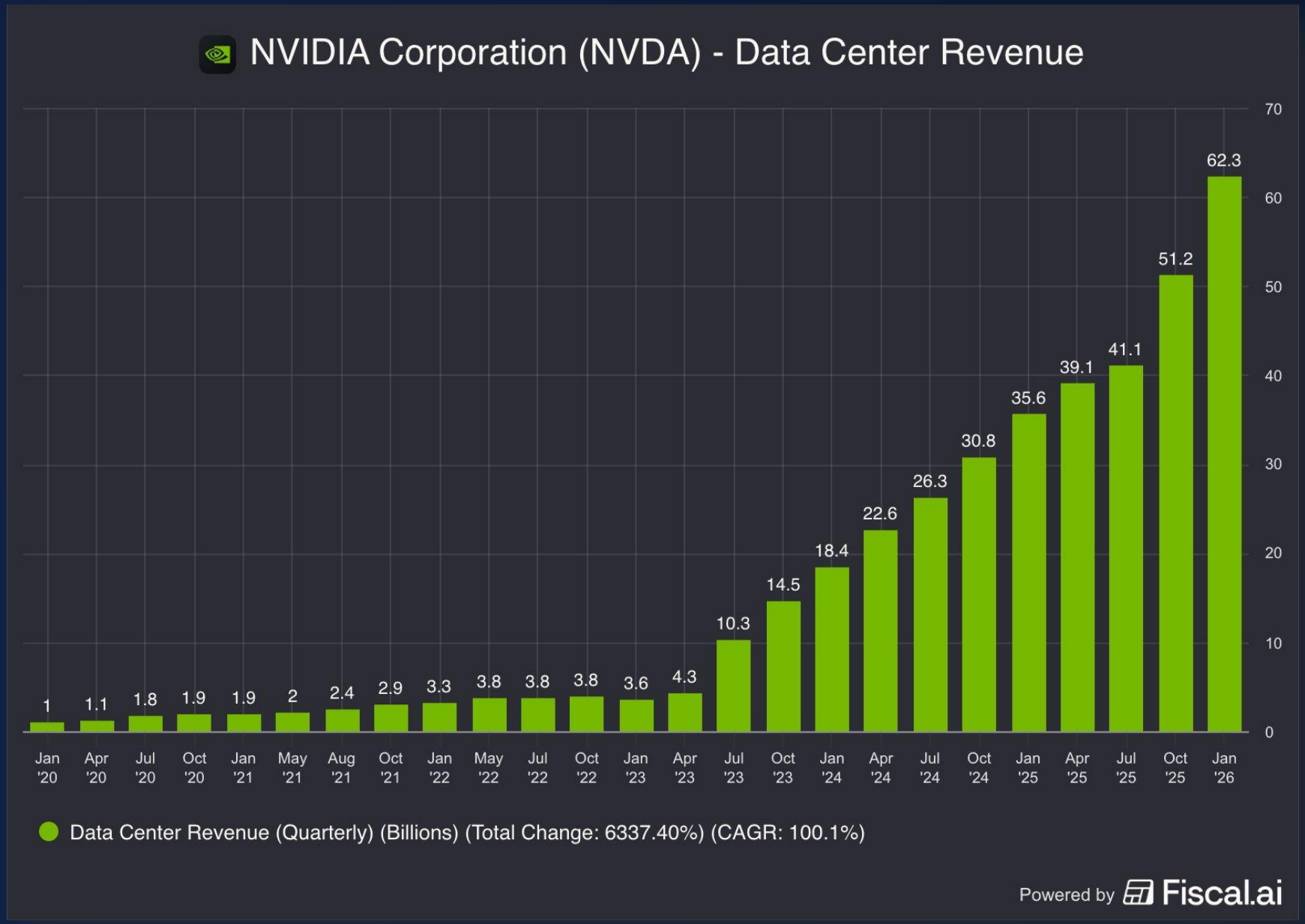
Entrenamiento
"Attention is all you need"
[GPUs]

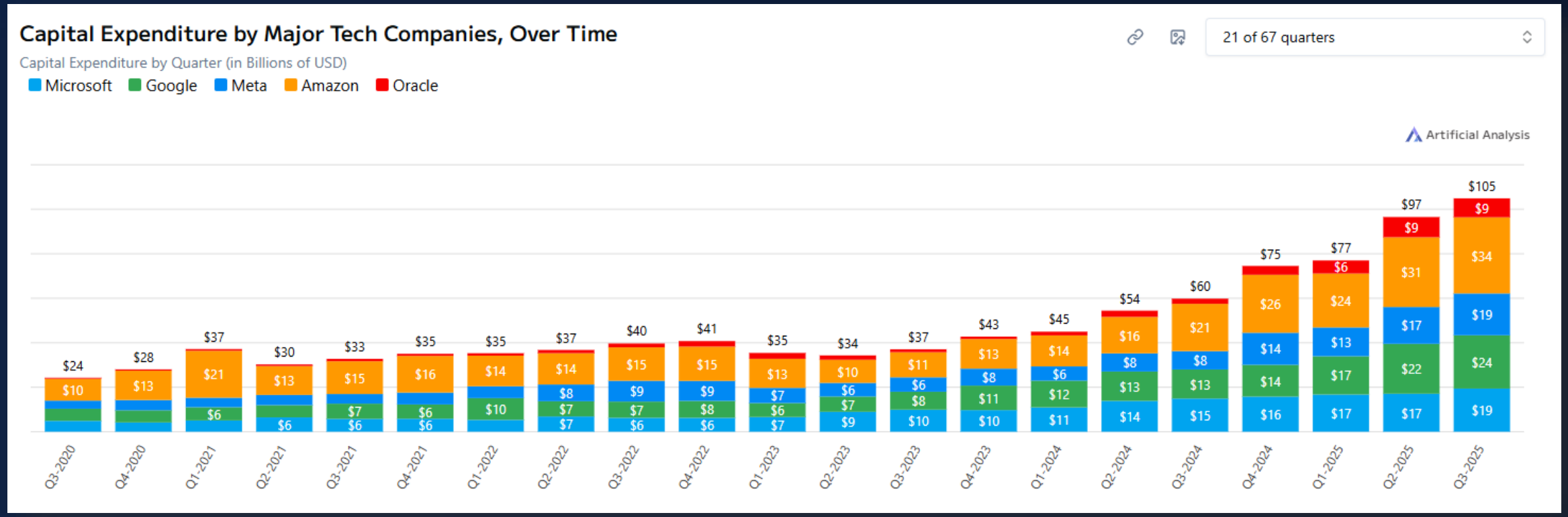
Fine-Tuning
[GPUs]

Large Language Model

Foundational Model

Nota: GPU ~= NVIDIA


NVIDIA Corporation (NVDA) - Data Center Revenue
70
60
50
40
30
20
10
0
1
1.1
1.8
1.9
1.9
2
2.4
2.9
3.3
3.8
3.8
3.8
3.6
4.3
10.3
14.5
18.4
22.6
26.3
30.8
35.6
39.1
41.1
51.2
62.3
Jan '20
Apr '20
Jul '20
Oct '20
Jan '21
May '21
Aug '21
Oct '21
Jan '22
May '22
Jul '22
Oct '22
Jan '23
Apr '23
Jul '23
Oct '23
Jan '24
Apr '24
Jul '24
Oct '24
Jan '25
Apr '25
Jul '25
Oct '25
Jan '26
Data Center Revenue (Quarterly) (Billions) (Total Change: 6337.40%) (CAGR: 100.1%)
Powered by Fiscal.ai



https://artificialanalysis.ai/

Scaling Laws for Neural Language Models

arXiv:2001.08361v1 [cs.LG] 23 Jan 2020

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess
OpenAI
bchess@openai.com

Rewon Child
OpenAI
rewon@openai.com

Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com

Jeffrey Wu
OpenAI
jeffwu@openai.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

We study empirical scaling laws for language model performance on the cross-entropy loss. The loss scales as a power-law with model size, dataset size, and the amount of compute used for training, with some trends spanning more than seven orders of magnitude. Other architectural details such as network width or depth have minimal effects within a wide range. Simple equations govern the dependence of overfitting on model/dataset size and the dependence of training speed on model size. These relationships allow us to determine the optimal allocation of a fixed compute budget. Larger models are significantly more sample-efficient, such that optimally compute-efficient training involves training very large models on a relatively modest amount of data and stopping significantly before convergence.

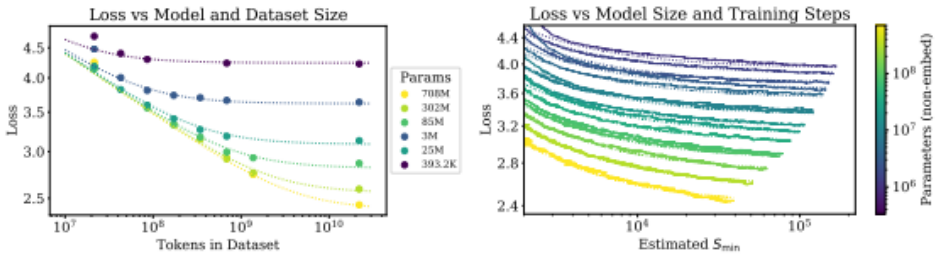


Figure 4 **Left**: The early-stopped test loss $L(N, D)$ varies predictably with the dataset size D and model size N according to Equation (1.5). **Right**: After an initial transient period, learning curves for all model sizes N can be fit with Equation (1.6), which is parameterized in terms of $S_{\min}$, the number of steps when training at large batch size (details in Section 5.1).

El modelo mejora en base a:

- Tamaño del modelo
- Cantidad de datos
- Entrenamiento realizado (pasos o cómputo)

https://arxiv.org/abs/2001.08361

Impacto para las personas y las organizaciones

Productividad + Mejora de Servicio

1

Productividad Personal

2

Productividad Organizacional

[Procesos]

3

Experiencia del cliente

[Apps & Servicios basados en GenAI]

4 GenAI en Desarrollo de Tecnología



Nov 2022

Inception

Fase de "Juguete Tecnológico". Generación de texto e imagen básica.

Chat



2024/2025

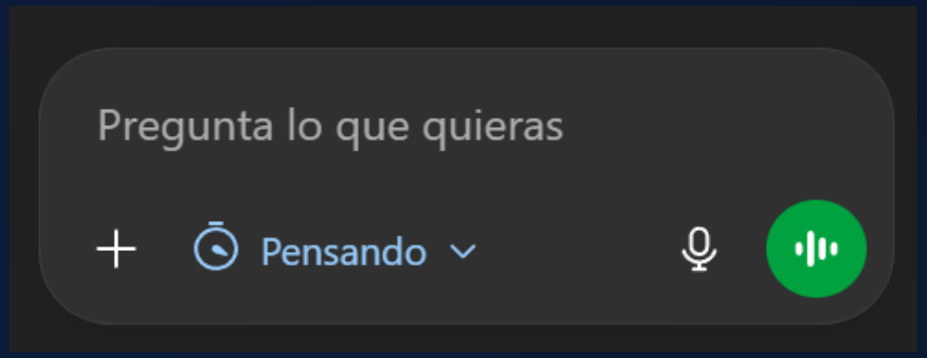
Razonamiento

Thinking Models. El sistema "piensa" y verifica antes de responder.

Chat e integración de LLMs en herramientas

Sistemas que “piensan” o “razonan”

- Comprenden el objetivo de la instrucción
- Descomponen un problema en subproblemas
- Planifican
- Realizan inferencia lógica y matemática
- Verifican
- Se corrigen al detectar inconsistencias
- Iteran hasta generar el resultado esperado



IA Generativa



Datos



Entrenamiento
"Attention is all you need"
[GPUs]

Foundational Model



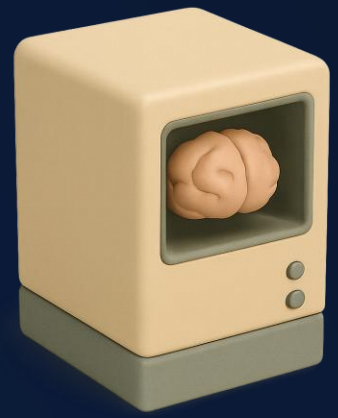
Fine-Tuning
[GPUs]





Inferencia
[GPUs]





Large Language Model
(Gemini, GPT, Claude, Grok, DeepSeek...)

GPU ~= NVIDIA



Nov 2022

Inception

Fase de "Juguete Tecnológico". Generación de texto e imagen básica.

Chat



2024/2025

Razonamiento

Thinking Models. El sistema "piensa" y verifica antes de responder.

Chat e integración de LLMs en herramientas



2025/2026

Agentes Autónomos

Sistemas que ven, escuchan, planifican y ejecutan acciones complejas.

Agentes que operan en forma autónoma, *con guía y supervisión*

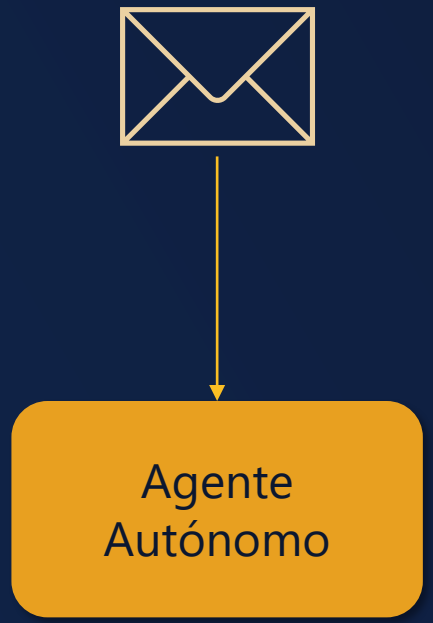
Agente autónomo

Sistema agéntico

Sistema que usa modelos avanzados (LLMs), herramientas y memoria para planificar y ejecutar acciones orientadas a un objetivo, ajustándose de manera iterativa según el contexto y los resultados.

Agentes autónomos

Ejemplo: Empresa de e-commerce



Reclamo de un cliente

Lee el contenido
Valida la información del reclamo: Tiempo, estado, etc
Historial del cliente
Recopila toda la información relevante sobre el reclamo
Establece un nivel de “criticidad”
Elabora una propuesta de respuesta
Decide a quién derivarlo

Nota: el Agente, planifica, evalúa, se conecta y dialoga con múltiples sistemas para obtener información, valida cada paso, hasta producir un resultado

Proceso del Agente Autónomo

- Entender el Objetivo
- Planificar las acciones
- Ejecutar múltiples pasos
- Interactuar con distintos sistemas
- Evaluar y Verificar
- Generar resultado

4 patrones de uso

a) Tareas de corta duración, guiadas por el usuario (ej DeepResearch)

b) Tareas disparadas automáticamente por un evento determinado (ej Correo reclamo)

c) Tareas de larga duración definidas por un usuario experto (ej Claude Cowork)

d) Agentes Autónomos 24x7 (ej OpenClaw)

Nuevo paradigma:
del chat que responde,
al agente autónomo que opera en base
a instrucciones para alcanzar un objetivo

Productividad Personal

Agentes Autónomos Guiados

Productividad Organizacional [procesos]

Agentes Autónomos

Apps & Servicios basados en GenAI

Integración / API

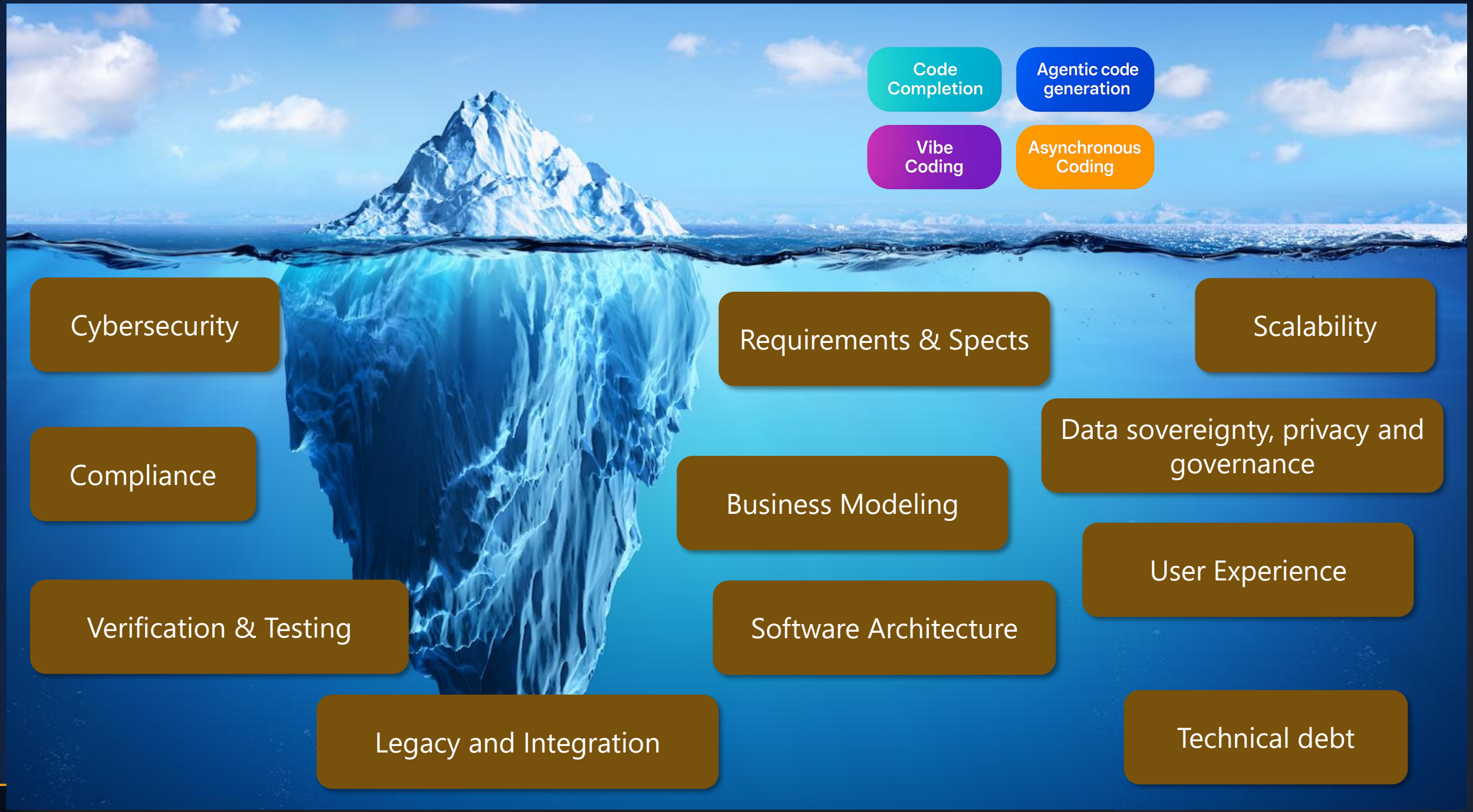
GenAI at Software Engineering

Asynchronous Coding Agents



Software Engineering

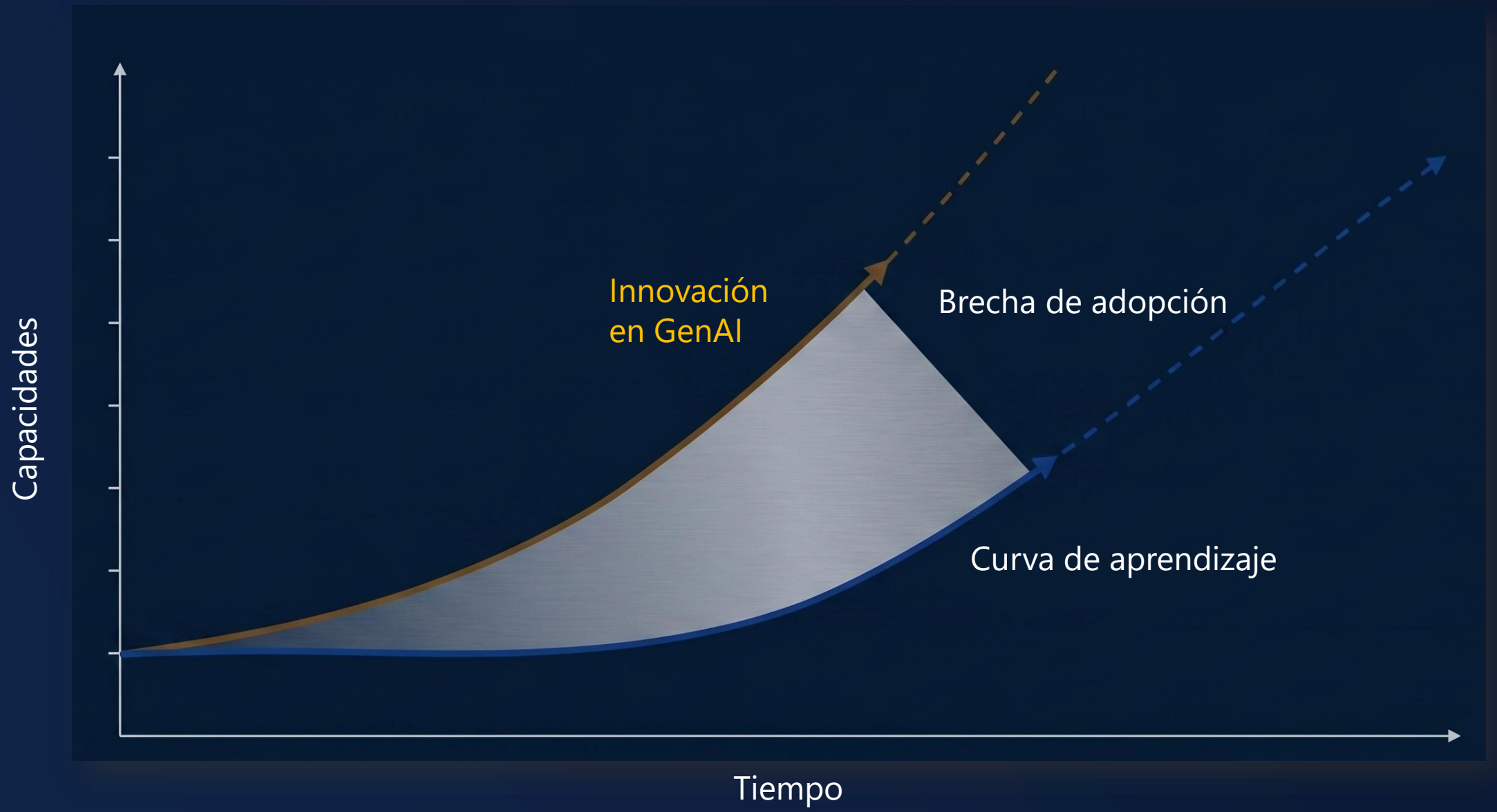
Construir software es mucho más que escribir código


Code Completion
Agentic code generation
Vibe Coding
Asynchronous Coding
Cybersecurity
Requirements & Spects
Scalability
Compliance
Data sovereignty, privacy and governance
Business Modeling
User Experience
Verification & Testing
Software Architecture
Legacy and Integration
Technical debt


Code Completion
Agentic code generation
Vibe Coding
Asynchronous Coding
Enterprise-Ready Software

Nueva alfabetización

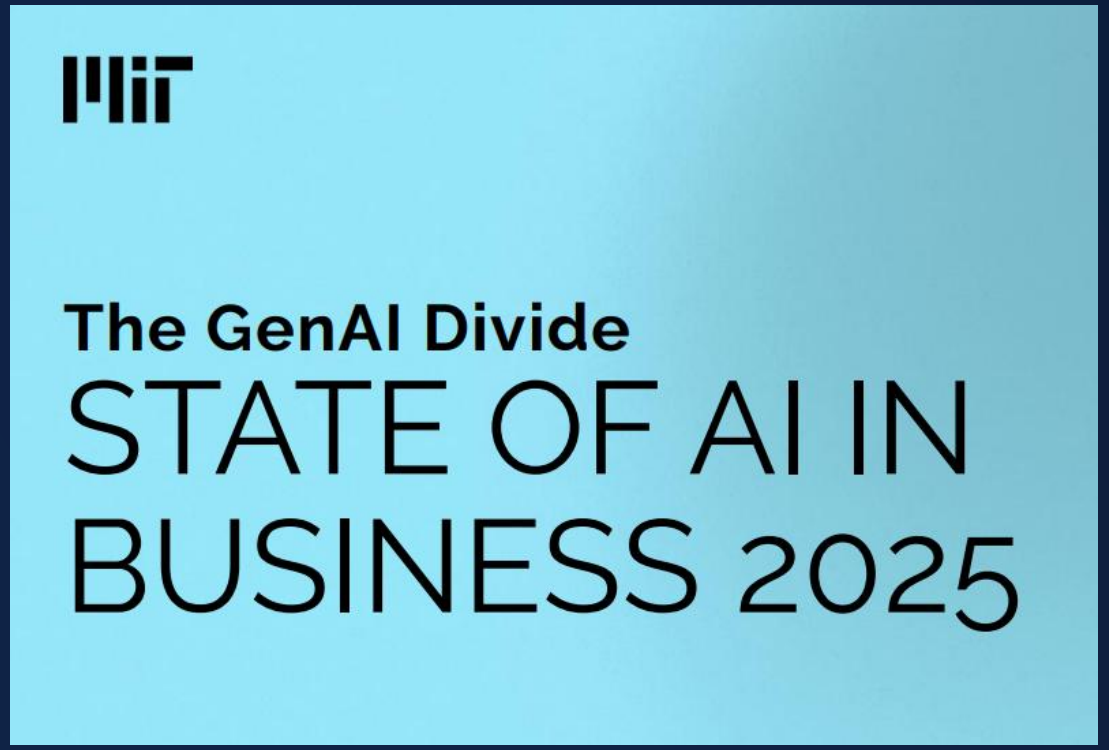
Para nosotros y nuestros equipos


Capacidades
Innovación en GenAI
Brecha de adopción
Curva de aprendizaje
Tiempo

Realidad vs Ruido

Difícil de equilibrar

MIT: 95%



NEWSLETTERS · CFO DAILY

MIT report: 95% of generative AI pilots at companies are failing



BY **SHERYL ESTRADA**
SENIOR WRITER AND AUTHOR OF CFO DAILY

August 18, 2025 at 6:54 AM EDT

https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

The Labor Market Effects of Generative Artificial Intelligence (Stanford University)

The study finds that **generative AI use at work can triple productivity** …



The Labor Market Effects of Generative Artificial Intelligence

36 Pages • Posted: 10 Apr 2025 • Last revised: 24 Oct 2025

Jonathan Hartley
Stanford University

Filip Jolevski
George Mason University - Mercatus Center; World Bank - Development Research Group (DECRG); Department of Economics, George Mason University

Vitor Melo
Clemson University

Brendan Moore
Stanford University, Department of Economics

Date Written: December 18, 2024

Realidad vs Ruido

Difícil de equilibrar

Solo podemos comprender el real impacto de los sistemas de IA Gen cuando entendemos con fundamentos cómo funcionan, sus capacidades y sus limitaciones.

Alcance y velocidad

Tecnológica
- Modelos/LLMs
- Infraestructura
- Datos
- Cómputo
- Empresas

Económica
- Productividad
- Mercados públicos de capital
- Inversión privada / VC
- Energía

Laboral
- Impacto en el empleo
- Automatización
- Nuevas habilidades

Organizacional
- Rediseño de procesos
- Nuevas formas de trabajo
- Gobernanza y gestión del cambio

Educativa
- Rediseño curricular
- Impacto para docentes y alumnos

Social y cultural
- Impacto social
- Acceso y desigualdad
- Sesgos

Política e institucional
- Regulación y derecho
- Políticas públicas

Geopolítica y Ambiental
- Competencia entre países
- Soberanía tecnológica
- Recursos e impacto ecológico

GenAI y la Intuición

https://bit.ly/ANIU_Mangarelli042026

Muchas gracias

mangarelli@ort.edu.uy

1

Productividad Personal

2

Productividad Organizacional [procesos]

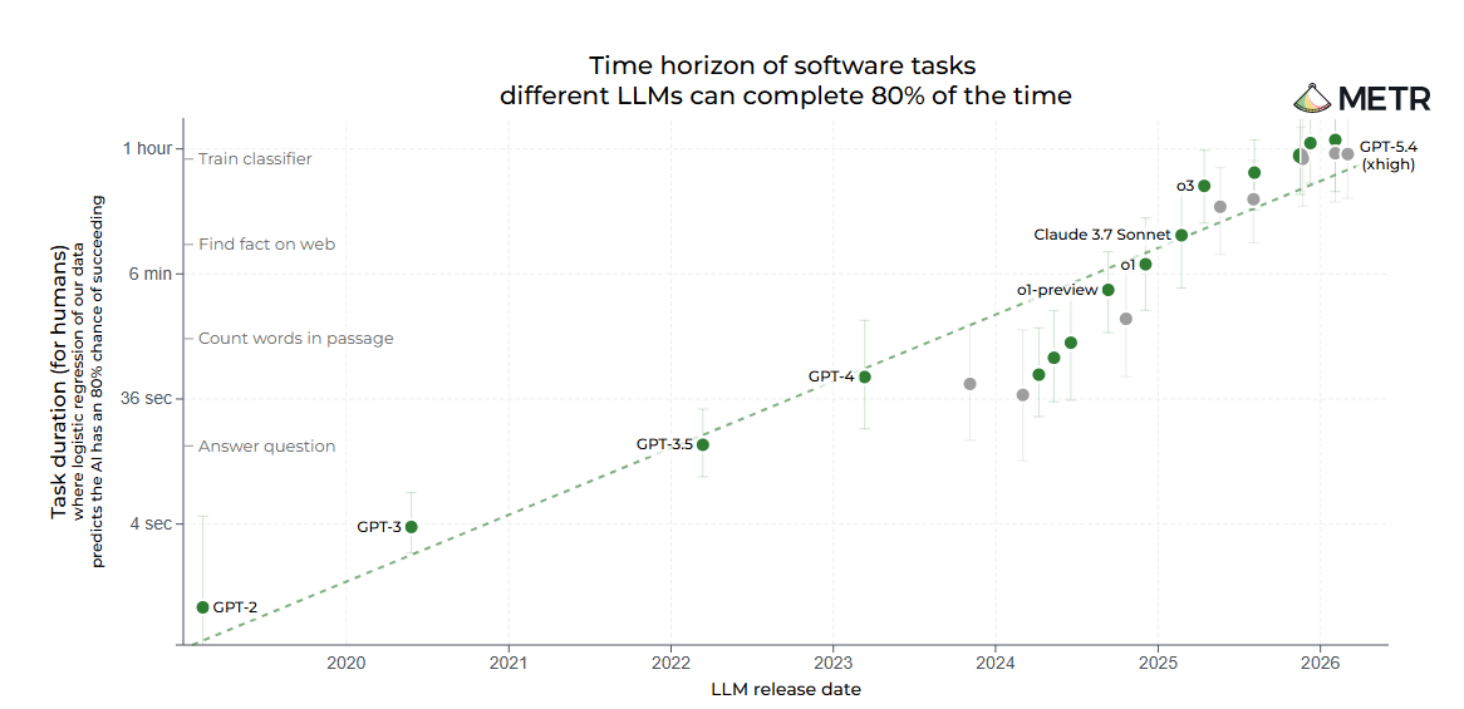
3

Apps & Servicios basados en GenAI

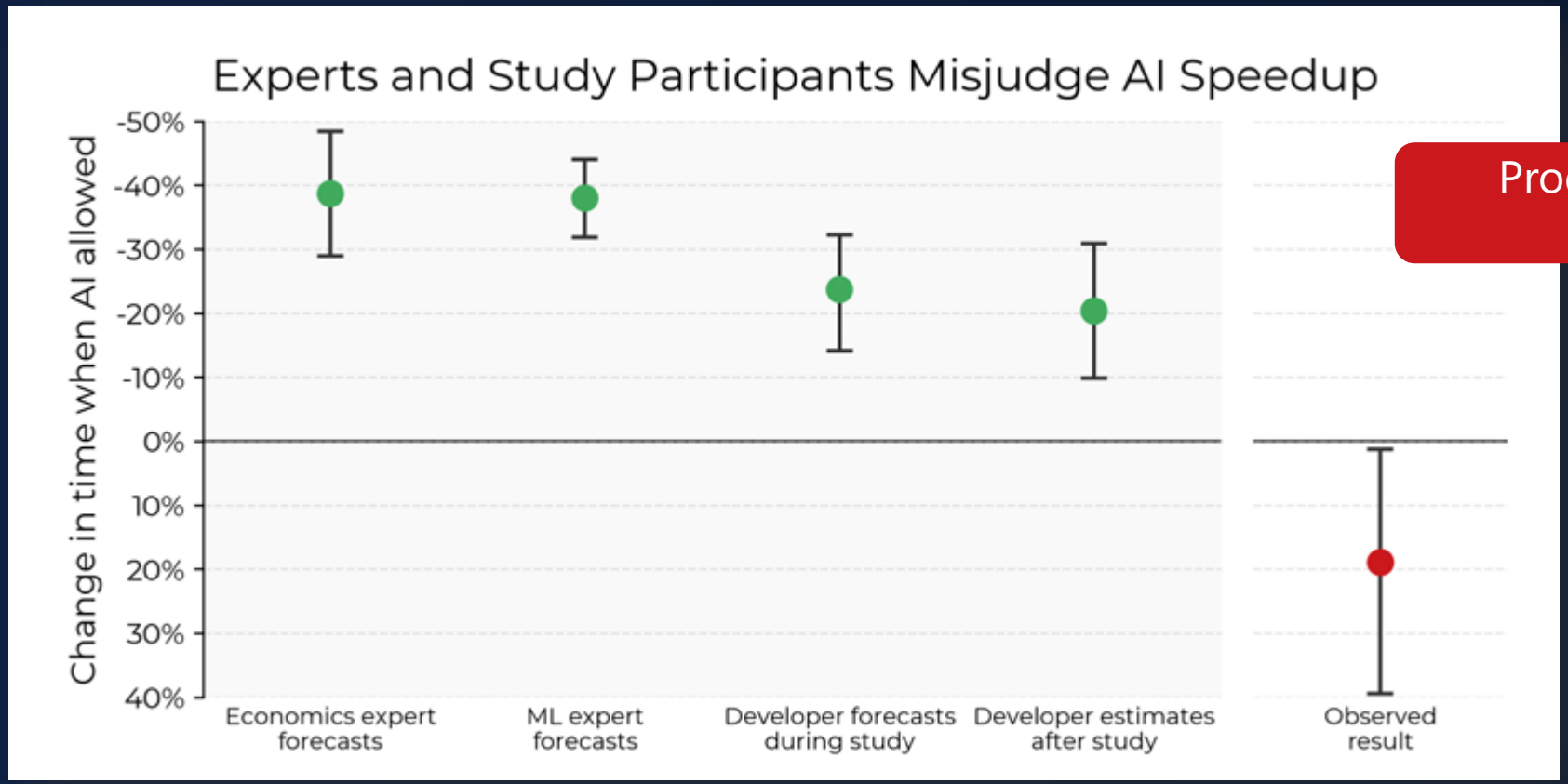
4 GenAI at Software Engineering

Measuring AI Ability to Complete Long Tasks

Tareas de hasta **54 minutos** se llevan adelante con un **80% de éxito**



Impact Experienced Open-Source Dev Productivity



Productividad
-19%

https://arxiv.org/abs/2507.09089

Desarrollo de software asistido por GenAI



End-User

Developer

Vibe Coding

Lovable.dev
Rork
Bolt.new
v0
Firebase Studio (Google)
Replit

Agentic code generation

Cursor
Windsurf
VS Code/GitHub Copilot
Antigravity

Code Completion

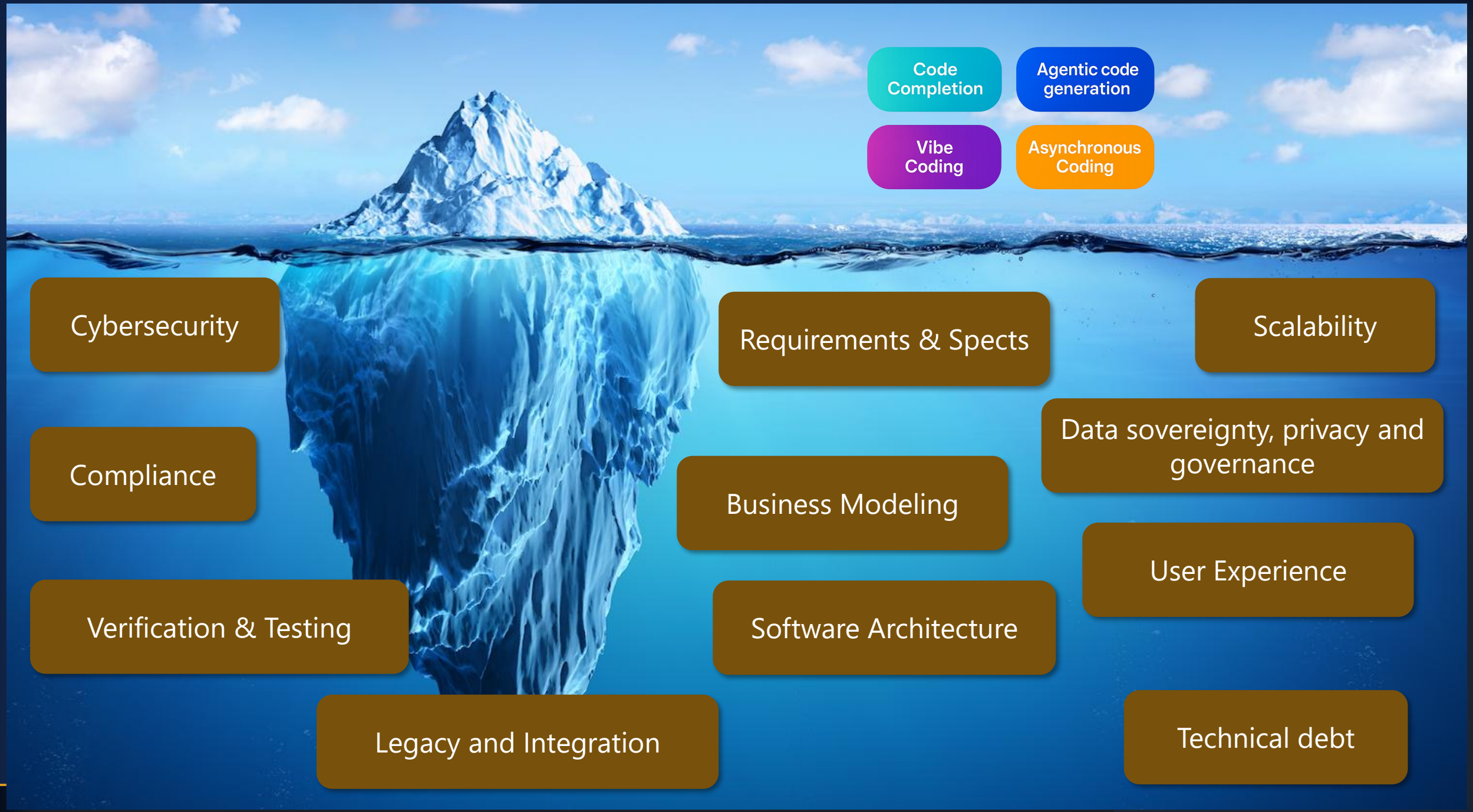
Asynchronous Coding

Codex (OpenAI)
Claude Code
Github Copilot Agent
Jules (Google)



Software Engineering

Construir software es mucho más que escribir código


Code Completion
Agentic code generation
Vibe Coding
Asynchronous Coding
Cybersecurity
Requirements & Spects
Scalability
Compliance
Data sovereignty, privacy and governance
Business Modeling
User Experience
Verification & Testing
Software Architecture
Legacy and Integration
Technical debt


Code Completion
Agentic code generation
Vibe Coding
Asynchronous Coding
Enterprise-Ready Software

Productividad Personal

Agentes Autónomos Guiados

Productividad Organizacional [procesos]

Agentes Autónomos

Apps & Servicios basados en GenAI

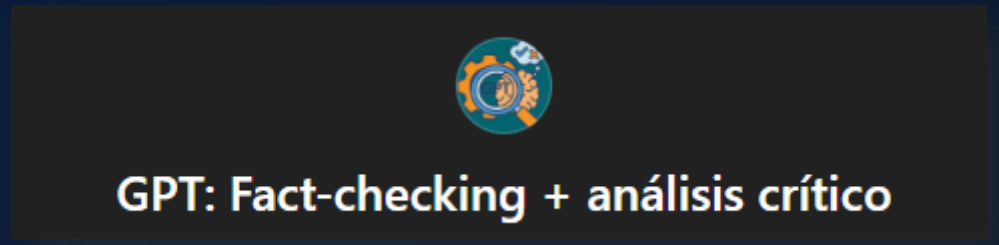
Integración / API

GenAI at Software Engineering

Asynchronous Coding Agents

Ej:
Fact checking y Verificación de conclusiones

- GPT, configurado para:
- Realizar fact checking
- Verificar evidencia-conclusión



https://bit.ly/factcheckingyverificacion