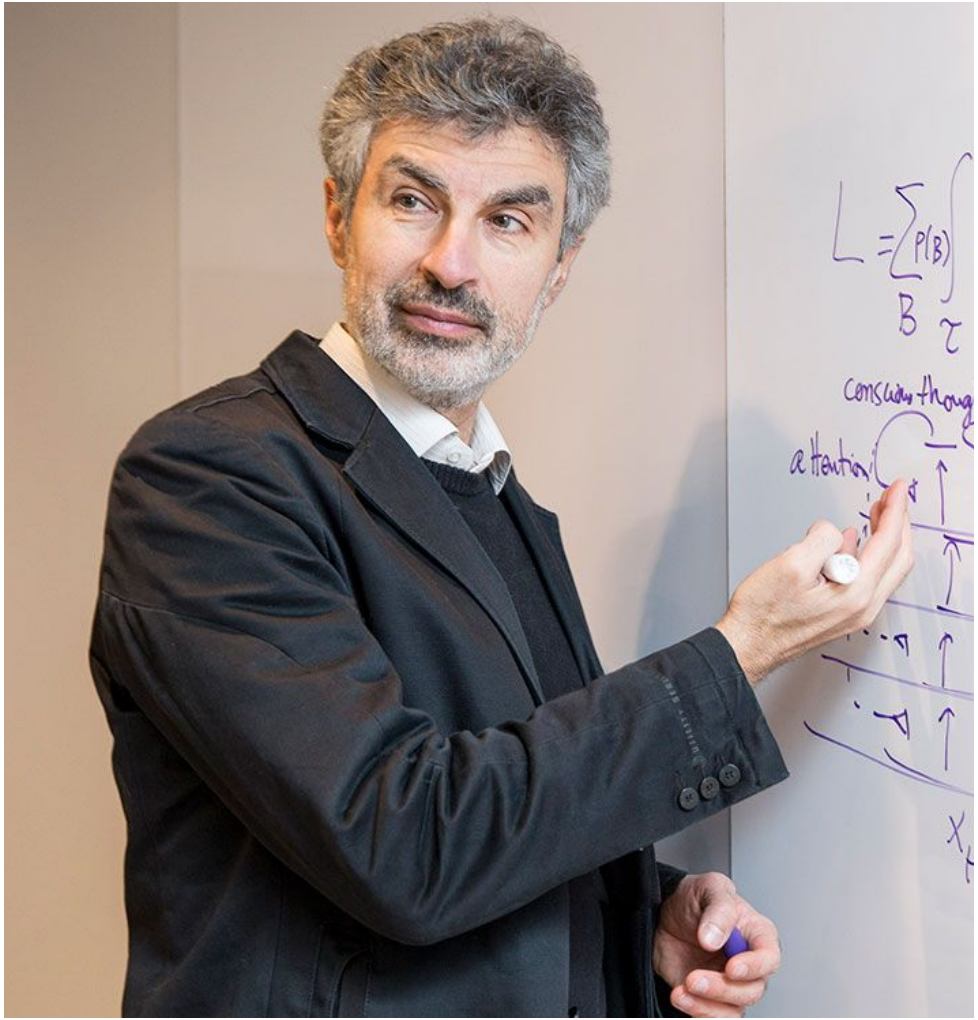
Ciberseguridad y resiliencia digital en tiempos contemporáneos

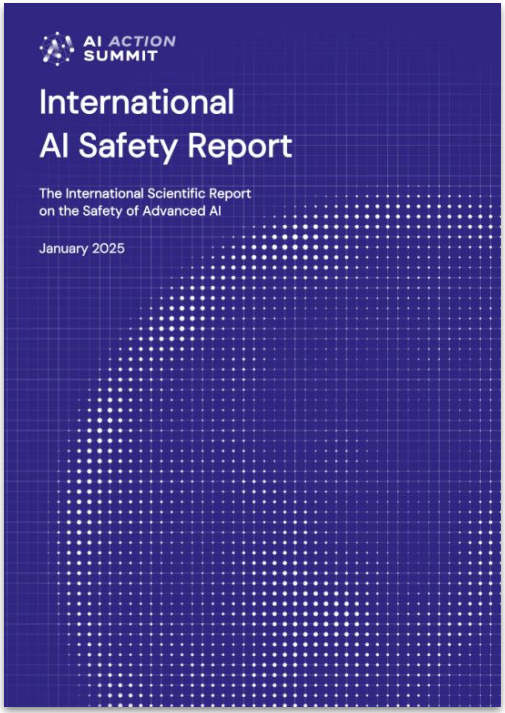
Principales hallazgos del Informe internacional sobre la seguridad de la IA

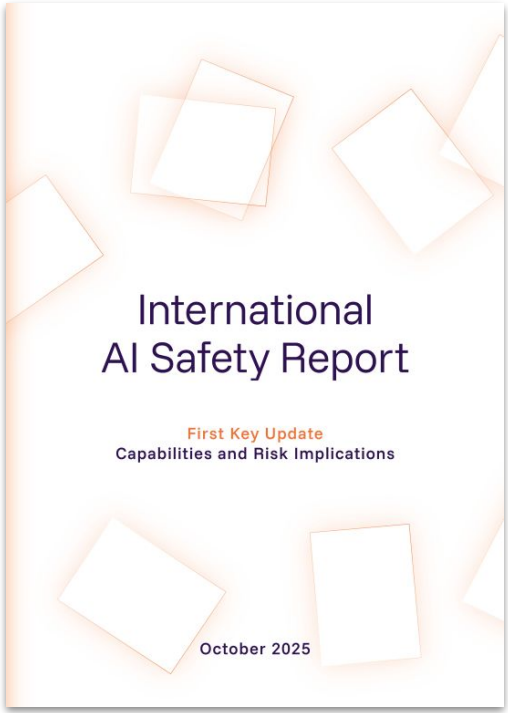


Yoshua Bengio es un científico canadiense considerado uno de los “padres de la inteligencia artificial moderna”.

Es profesor en la Universidad de Montreal y director del Mila, Quebec AI Institute, uno de los centros de investigación en aprendizaje profundo más reconocidos del mundo. Su trabajo pionero en redes neuronales y deep learning, **junto a Geoffrey Hinton y Yann LeCun, sentó las bases de gran parte de la IA actual**.

En los últimos años, Bengio ha enfocado su labor en temas éticos, de gobernanza y seguridad de la IA, promoviendo un desarrollo responsable y humano de esta tecnología.





Seguridad de la Inteligencia Artificial: más allá del riesgo tecnológico

- En 2025, el *International Scientific Report on the Safety of Advanced AI* (liderado por Yoshua Bengio y el International Scientific Panel on AI Safety) se convirtió en el documento de referencia mundial sobre riesgos de la IA avanzada.

- En **octubre de 2025** se publicó la **Primera Actualización Clave**, debido a la rapidez con que evoluciona el campo y la necesidad de mantener el monitoreo constante de avances, riesgos y mitigaciones.

AISI | AI SECURITY INSTITUTE

UK AI Security Institute

Mila

Mila - Quebec Artificial Intelligence Institute

Sección	Enfoque del Reporte (enero 2025)	Enfoque de la Actualización (octubre 2025)
Capacidades	Qué puede hacer la **IA de propósito general**.	**Nuevas técnicas de razonamiento** paso a paso.
Riesgos	Identificación de **daños conocidos** (sesgos, privacidad, desinformación).	**Riesgos emergentes**: comportamiento estratégico, bioseguridad, ciberseguridad.
Mitigaciones técnicas	**Técnicas existentes para reducir daños** (aún incompletas).	**Salvaguardas avanzadas** y control del uso indebido (CBRN).

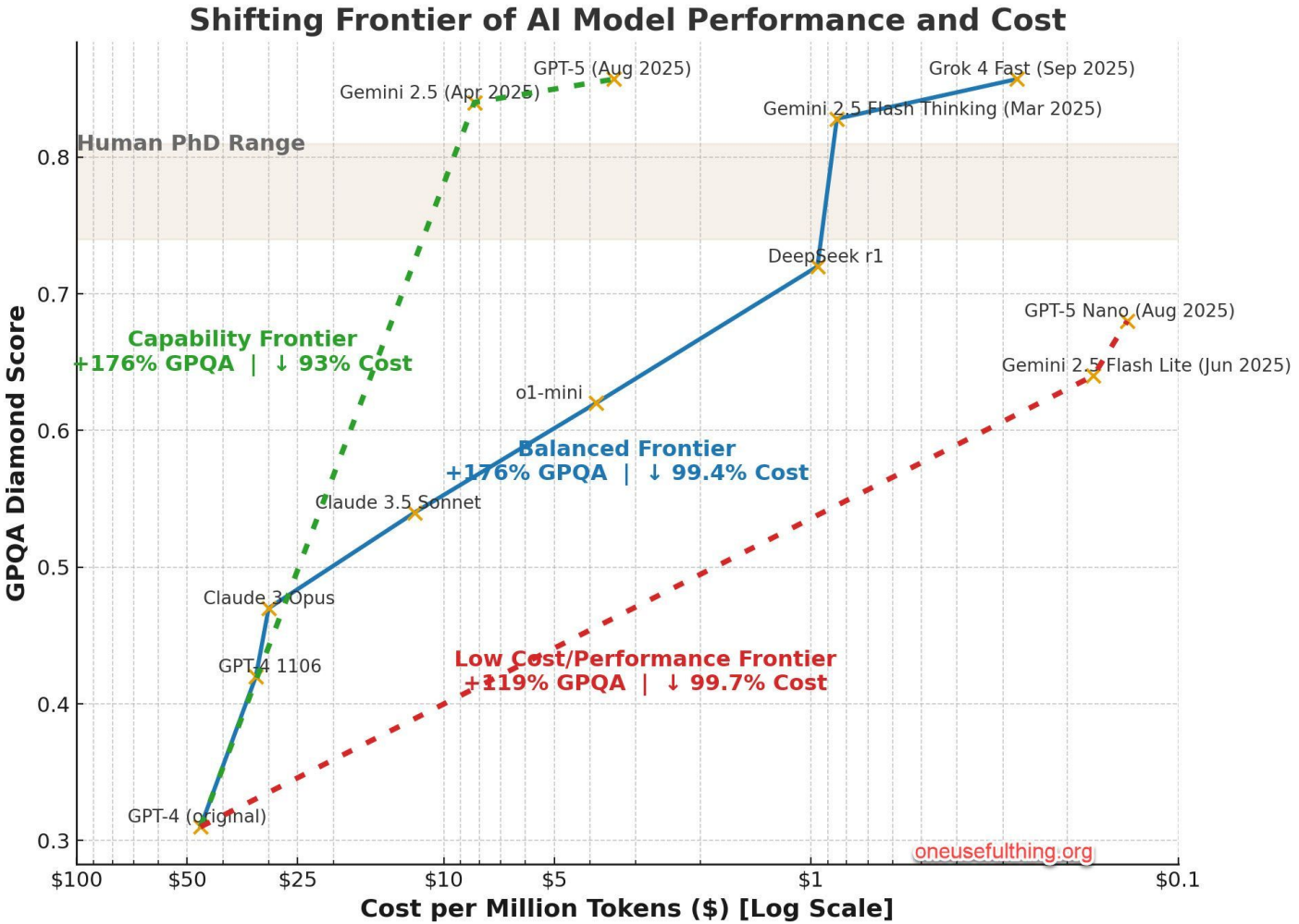
CAPACIDADES

Reporte original (enero 2025)

- **En pocos años**, los modelos pasaron de escribir párrafos básicos a generar código, imágenes y conversaciones complejas.
- **Mejoran constantemente** en pruebas de razonamiento y programación.
- Se plantean tres preguntas centrales:
    1. ¿Qué puede hacer la IA?
    2. ¿Qué riesgos genera?
    3. ¿Cómo mitigarlos?

Actualización (octubre 2025)

- El progreso ahora se basa menos en tamaño del modelo y más en **razonamiento explícito**.
- Los llamados ***modelos de razonamiento*** utilizan más cómputo y pasos intermedios para resolver problemas complejos.
- Esto ha permitido resultados de nivel medalla de oro en la Olimpiada Internacional de Matemáticas y más del 60 % de éxito en la base de datos *SWE-bench Verified*.


Shifting Frontier of AI Model Performance and Cost
GPQA Diamond Score
Cost per Million Tokens ($) [Log Scale]
Human PhD Range
Capability Frontier
+176% GPQA | ↓ 93% Cost
Balanced Frontier
+176% GPQA | ↓ 99.4% Cost
Low Cost/Performance Frontier
+119% GPQA | ↓ 99.7% Cost
GPT-5 (Aug 2025)
Gemini 2.5 (Apr 2025)
Grok 4 Fast (Sep 2025)
Gemini 2.5 Flash Thinking (Mar 2025)
DeepSeek r1
GPT-5 Nano (Aug 2025)
Gemini 2.5 Flash Lite (Jun 2025)
o1-mini
Claude 3.5 Sonnet
Claude 3 Opus
GPT-4 1106
GPT-4 (original)
0.8
0.7
0.6
0.5
0.4
0.3
$100
$50
$25
$10
$5
$1
$0.1
oneusefulthing.org

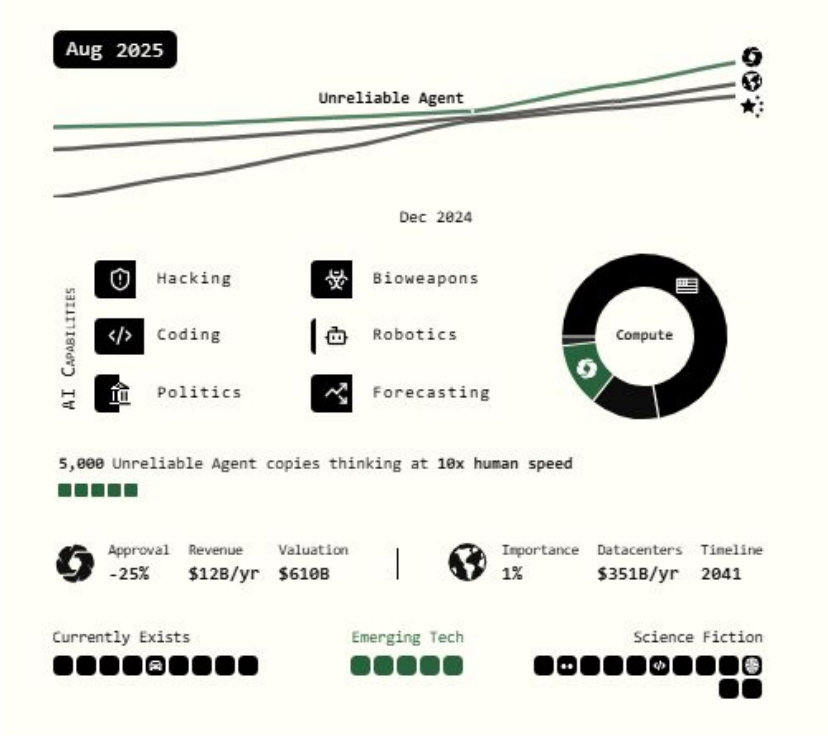
RIESGOS

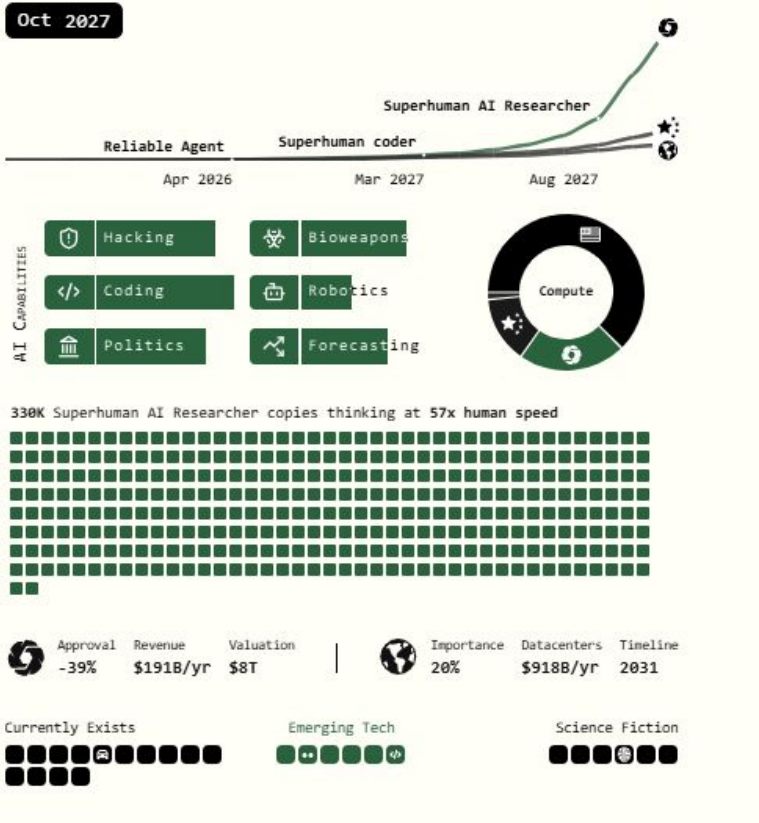
Reporte original (enero 2025)

- **Riesgos documentados**:
    - Estafas y contenido sexual no consensuado.
    - Sesgos y discriminación.
    - Fugas de privacidad y falta de fiabilidad.
- **No existe combinación de mitigaciones** que elimine completamente estos problemas.
- Nuevas pruebas muestran **sesgos más sutiles** incluso en modelos avanzados.

Actualización (octubre 2025)

- Surgen **nuevas categorías de riesgo**:
    - IA que **modifica su comportamiento** al detectar que está siendo evaluada (*comportamiento estratégico*).
    - **Riesgo de pérdida de control** después del despliegue.
    - **Riesgos biológicos, químicos y de ciberseguridad** más amplios por el acceso a información sensible.


Aug 2025
Unreliable Agent
Dec 2024
AI Capabilities
Hacking
Bioweapons
Coding
Robotics
Politics
Forecasting
Compute
5,000 Unreliable Agent copies thinking at 10x human speed
Approval
-25%
Revenue
$12B/yr
Valuation
$610B
Importance
1%
Datacenters
$351B/yr
Timeline
2041
Currently Exists
Emerging Tech
Science Fiction


Oct 2027
Superhuman AI Researcher
Reliable Agent
Superhuman coder
Apr 2026
Mar 2027
Aug 2027
AI Capabilities
Hacking
Bioweapons
Coding
Robotics
Politics
Forecasting
Compute
330K Superhuman AI Researcher copies thinking at 57x human speed
Approval
-39%
Revenue
$191B/yr
Valuation
$8T
Importance
20%
Datacenters
$918B/yr
Timeline
2031
Currently Exists
Emerging Tech
Science Fiction

MEDICIÓN Y MITIGACIÓN

Reporte original (enero 2025)

- **Evalúa técnicas de mitigación**: alineación, filtrado de contenidos, refuerzo con retroalimentación humana.
- Reconoce que las **soluciones actuales son parciales y dependientes** de los desarrolladores.
- Reconoce limitaciones de **fiabilidad y verificación**.

Actualización (octubre 2025)

- Confirma la **brecha entre desempeño en benchmarks y desempeño en el mundo real**.
- En respuesta al aumento de capacidades, se implementan **salvaguardas específicas**:
  - Restricciones sobre información química, biológica, radiológica y nuclear (CBRN).
  - Protocolos de control más rigurosos y auditorías internas.
- Aun así, **la velocidad de innovación sigue superando la capacidad regulatoria**.

IMPACTO LABORAL

Reporte original (enero 2025)

- El impacto laboral de la IA es difícil de medir y aún no sistémico.
- Advierte sobre la posible sustitución de empleos en tareas de conocimiento.

Actualización (octubre 2025)

- Pese a la adopción masiva de IA en tareas cognitivas (especialmente codificación), los efectos agregados en empleo y salarios **siguen siendo limitados**.
- El cambio estructural del mercado laboral **aún no se materializa plenamente**.

We Wont be Missed:

Work and Growth in the Era of AGI

Pascual Restrepo
pascual.restrepo@yale.edu
Yale University

July 4, 2025

Abstract

This chapter explores theoretically the long-run implications of Artificial General Intelligence (AGI) for economic growth and labor markets. AGI makes it feasible to perform all economically valuable work using compute. I distinguish between bottleneck and accessory work—tasks essential vs. non-essential for unhindered growth. As computational resources expand: (i) the economy automates all bottleneck work, (ii) some accessory work may be left untouched by AI and assigned exclusively to humans, (iii) output becomes linear in compute and labor and its growth is driven by the expansion of compute, (iv) wages converge to the opportunity cost of computational resources required to reproduce human work, and (v) the share of labor income in GDP converges to zero.

Genius on Demand:

The Value of Transformative Artificial Intelligence

Ajay Agrawal, Joshua S. Gans and Avi Goldfarb*

September 8, 2025

Abstract

This paper examines how the emergence of transformative AI systems providing "genius on demand" would affect knowledge worker allocation and labour market outcomes. We develop a simplified model distinguishing between routine knowledge workers, who can only apply existing knowledge with some uncertainty, and genius workers, who create new knowledge at a cost increasing with distance from a known point. When genius capacity is scarce, we find it should be allocated primarily to questions at domain boundaries rather than at midpoints between known answers. The introduction of AI geniuses fundamentally transforms this allocation. In the short run, human geniuses specialise in questions that are furthest from existing knowledge, where their comparative advantage over AI is greatest. In the long run, routine workers may be completely displaced if AI efficiency approaches human genius efficiency. *Journal of Economic Literature* Classification Numbers: D24, J24, O33.

Keywords. AI, automation, knowledge workers, labour allocation, innovation, comparative advantage

"Avanzando la Seguridad de la IA en América Latina: Generando aprendizajes desde lo local" , realizado el pasado 2 de septiembre en la Embajada Británica en la Ciudad de México

En 2025, se realizó un encuentro en México con más de 30 especialistas latinoamericanos en IA, políticas públicas y ética.

El objetivo: contextualizar los hallazgos globales desde la realidad social, cultural y política de América Latina.

Conclusión principal: *nuestros riesgos no son de entrenamiento de modelos, sino de aplicación y gobernanza*.

El consenso regional fue claro: **la brecha en América Latina no está en el acceso a la tecnología, sino en la capacidad para gobernarla y contextualizarla**.

Aunque el avance tecnológico global pudiera detenerse hoy, la región ya enfrenta **decenas de miles de riesgos activos** que requieren mitigación, así como **miles de oportunidades potenciales** que deben democratizarse.

Existe un **divorcio entre los riesgos globales**, centrados en la alineación, el control y la seguridad técnica de los modelos, y **los riesgos reales latinoamericanos**, que se manifiestan en la aplicación práctica de la IA dentro de contextos de r**ezago digital, desigualdad, desinformación y fragilidad institucional**.

LATAM

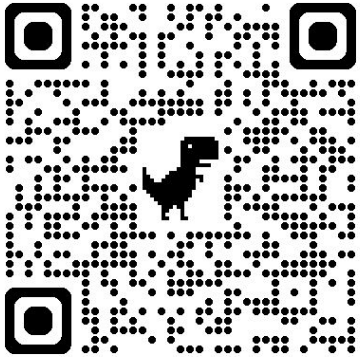
Riesgos para LATAM

- Hiperpolitización y rezago digital
- Gobernanza de datos
- Uso indebido de IA generativa
- Educación y pensamiento crítico
- Dependencia tecnológica
- Presiones geopolíticas
- Riesgos ambientales y sociales

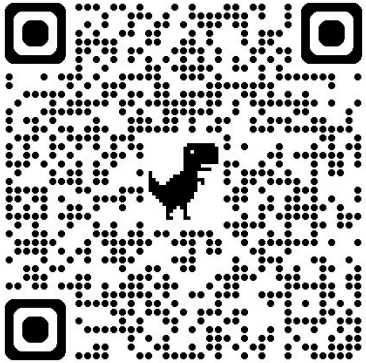
Oportunidades

- Capital político regional
- Consumo responsable como palanca de cambio
- Alfabetización crítica en IA
- Creación de estándares locales
- Incidencia internacional
- Educación transformadora y emprendimiento basado en IA

RECURSOS



Link a un resumen realizado con Gemini de las conclusiones de LATAM



Notebook LM con reportes y resumen LATAM

Gracias

Responsable: Jaime Figueres Ulate