COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI ICADE CIHS

Fundamentos de IA generativa

Álvaro López (Comillas – ICAI)

comillas.edu

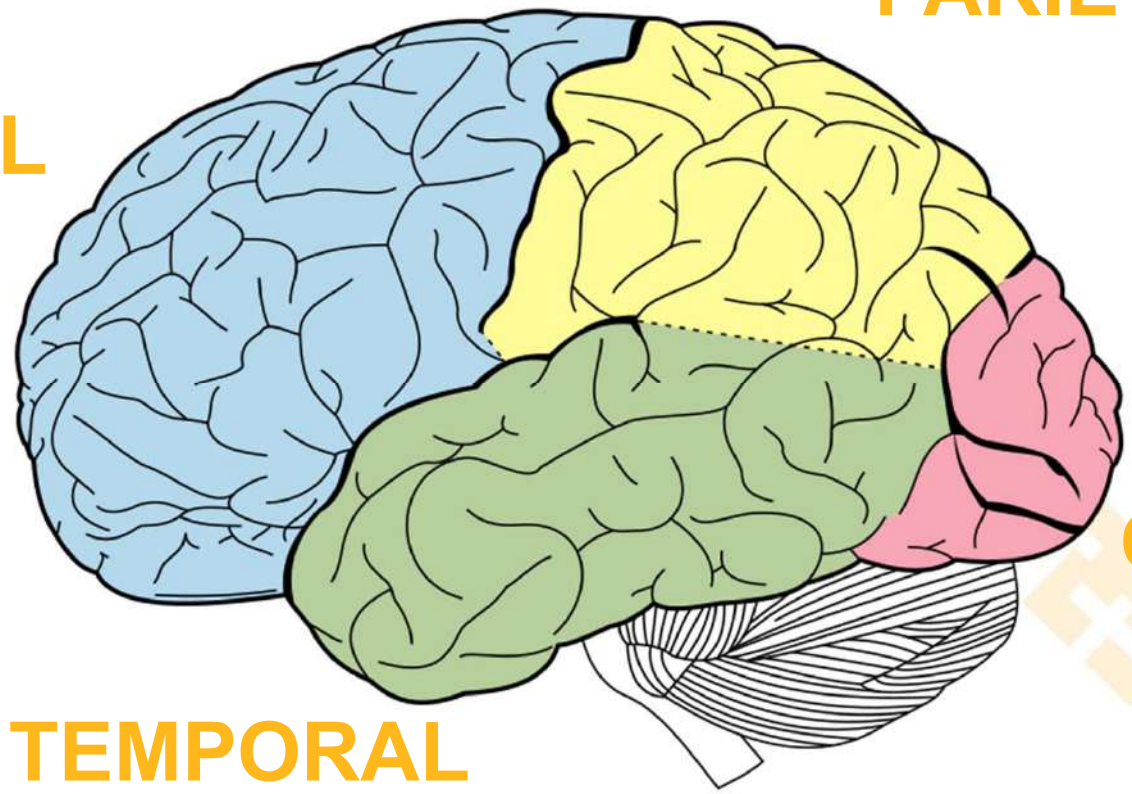
¿En qué punto está la inteligencia de las máquinas?

- **¿Qué referencia utilizamos para evaluarlo?**
- Se suele utilizar la **inteligencia humana**.
- Tiene sentido por la **fuerte influencia** sobre la artificial que ha tenido
- OJO: **no** son **fáciles de comparar**

comillas.edu

Ilustración creada con IA (DALL-E 3)

Sobre la inteligencia humana



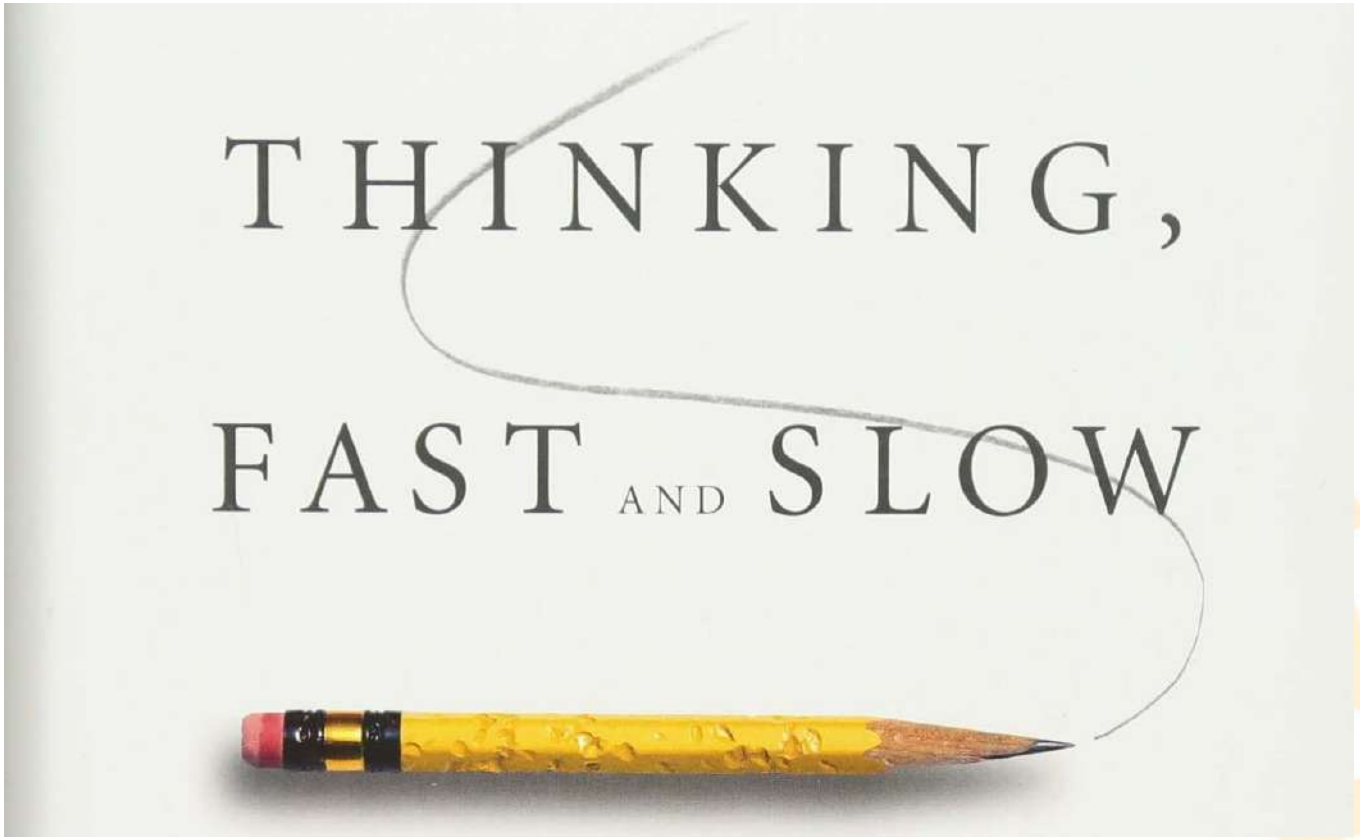
Un modelo muy útil

SISTEMA 1

Inconsciente
Rápido

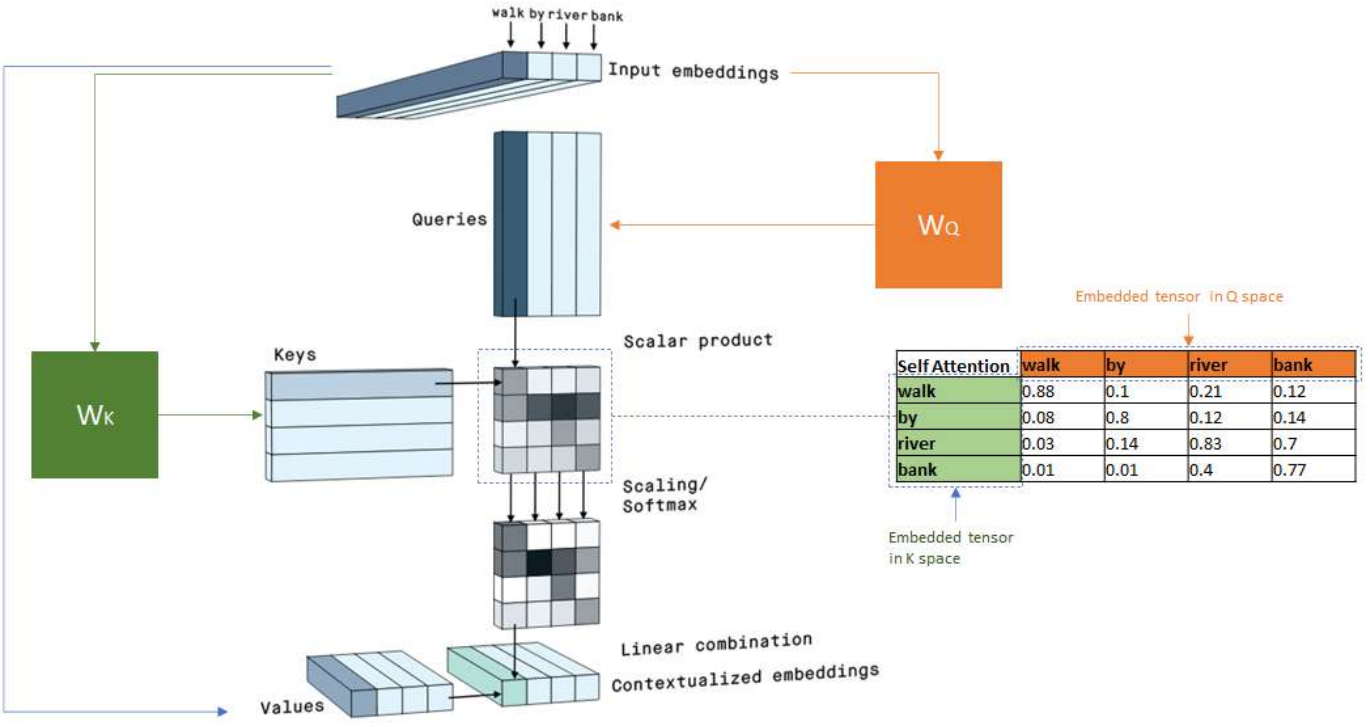
SISTEMA 2

Consciente, sofisticado
Lento



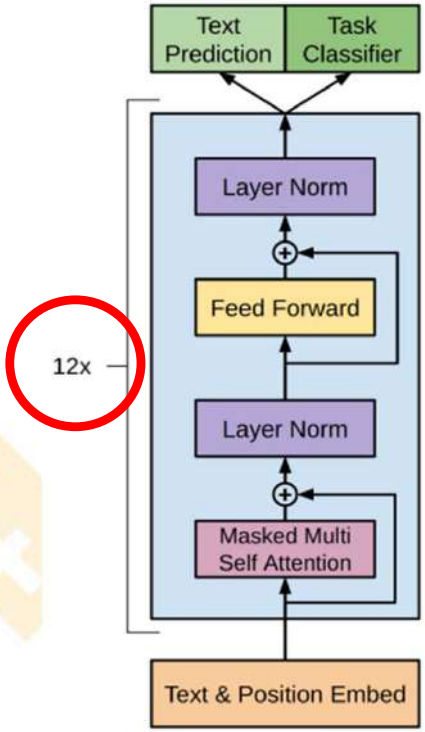
Sobre la “nueva” IA (LLMs)
Transformer

ATENCIÓN



Self Attention	walk	by	river	bank
walk	0.88	0.1	0.21	0.12
by	0.08	0.8	0.12	0.14
river	0.03	0.14	0.83	0.7
bank	0.01	0.01	0.4	0.77

GPT



Fuentes:

* https://stats.stackexchange.com/users/105137/mon), What exactly are keys, queries, and values in attention mechanisms?
* Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Estrategia de Pre-entrenamiento

MODELADO DEL LENGUAGE

- Accedemos a un texto sin etiquetar

> La capacidad de crear nuevos ejemplos parecidos a los ejemplos con los que se entrenó define a la nueva IA generativa

- Le “tapamos” la última palabra

> La capacidad de crear nuevos ejemplos parecidos a los ejemplos con los que se entrenó define a la nueva IA [redacted]

- Hacemos que la red neuronal aprenda a **proponer la siguiente palabra** dada una frase con cierto sentido.

- Manejo sobrehumano del **lenguaje**

Sólo el lenguaje… que no es poco

Symbols and mental programs: a hypothesis about human singularity

Stanislas Dehaene, [1,2,*] Fosca Al Roumi, [1] Yair Lakretz, [1] Samuel Planton, [1] and Mathias Sablé-Meyer [1]

- El lenguaje **no comparte circuitos neuronales** con los núcleos dedicados al pensamiento matemático, musical, geométrico, etc.
- **Pero nos permite explicar** cómo realizamos esos pensamientos.
- Grandes bases de datos de lenguaje humano contienen, **de forma indirecta**, algoritmos, modelos, abstracciones, etc.

Pero, entonces… ¿le enseñamos a hablar y ya está?

RLHF: Modelo **busca una recompensa** diferida ➔ no sólo emite palabras en base a las palabras previas

2018 - 2021

PRE-ENTRENAMIENTO

Enseñamos al modelo a manejar el lenguaje

GPT1 a GPT3.5

2022

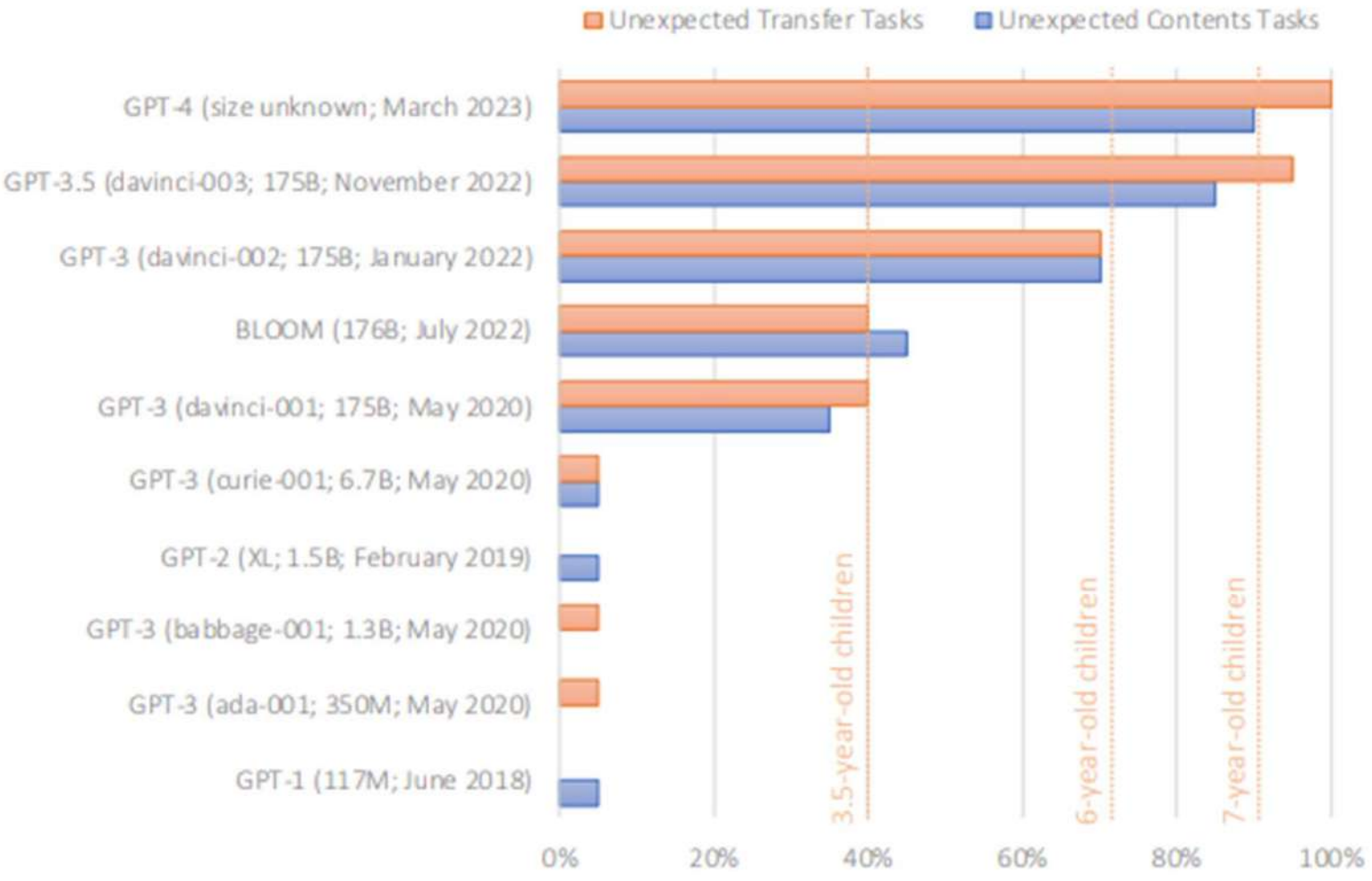
SEGUIMIENTO DE INSTRUCCIONES

Alineamiento. Enseñamos al modelo a satisfacer a su interlocutor

(Instruct GPT) Chat GPT

De manejar el lenguaje… a algo más

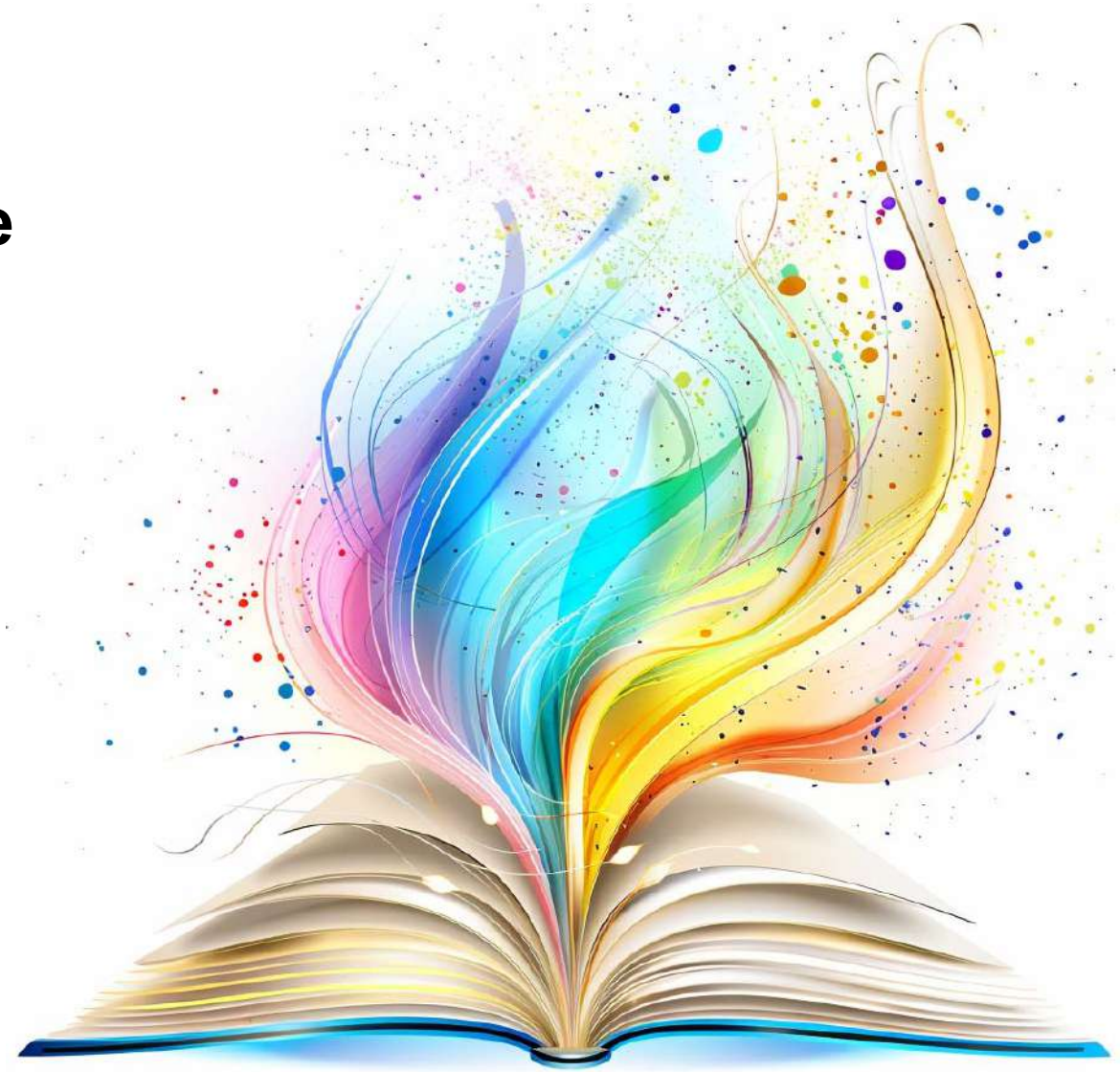
Teoría de la mente: capacidad de imputar estados mentales internos a otros estados. Única en el ser humano.



Fuente: https://arxiv.org/abs/2302.02083

GPT grande… LLM

- Interacción en **lenguaje natural**
- Resuelve **tareas de diversa índole**
- Está teniendo **gran impacto** en distintos sectores

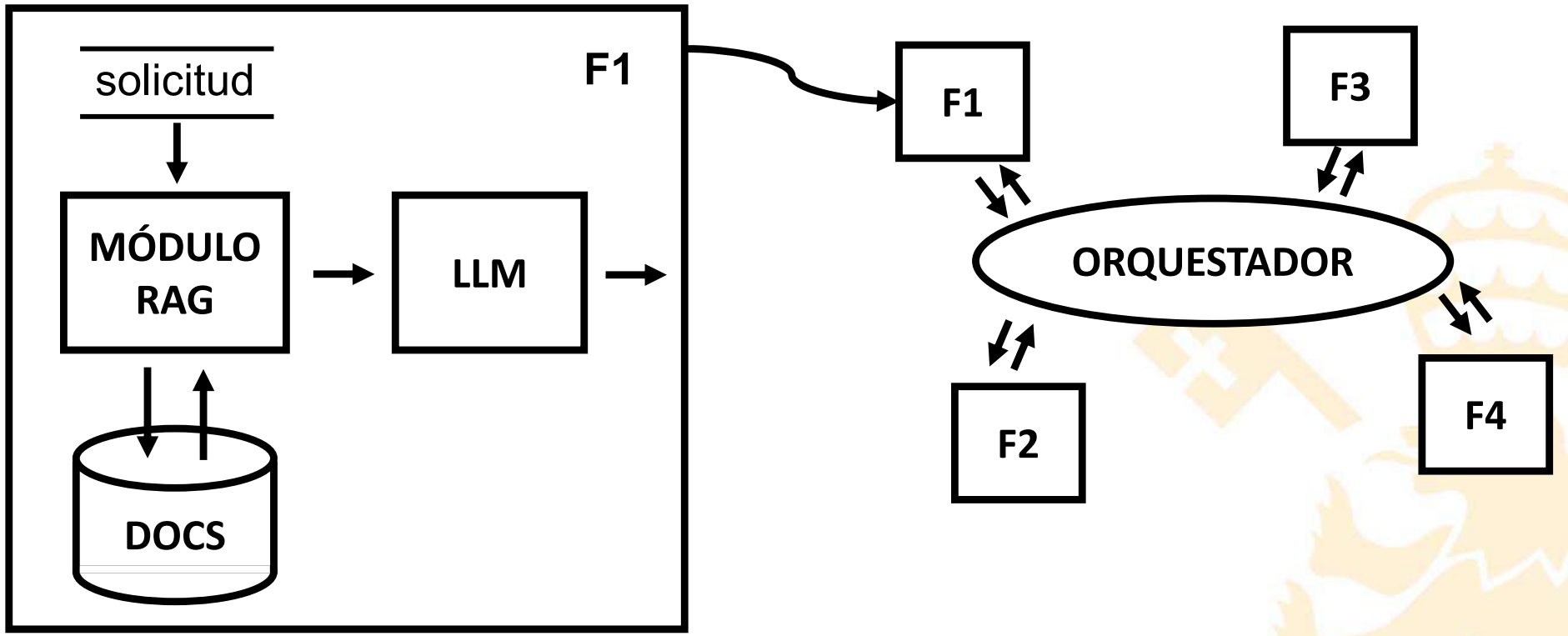


comillas.edu

Ilustración creada con IA (Midjourney)

Aplicaciones basadas en LLMs

- Están en el **núcleo de la mayoría de las aplicaciones** que nos están sorprendiendo desde finales de 2022.
- **Numerosas técnicas** para mejorar rendimiento aplicación



Parece magia, pero…
¿Tiene limitaciones?

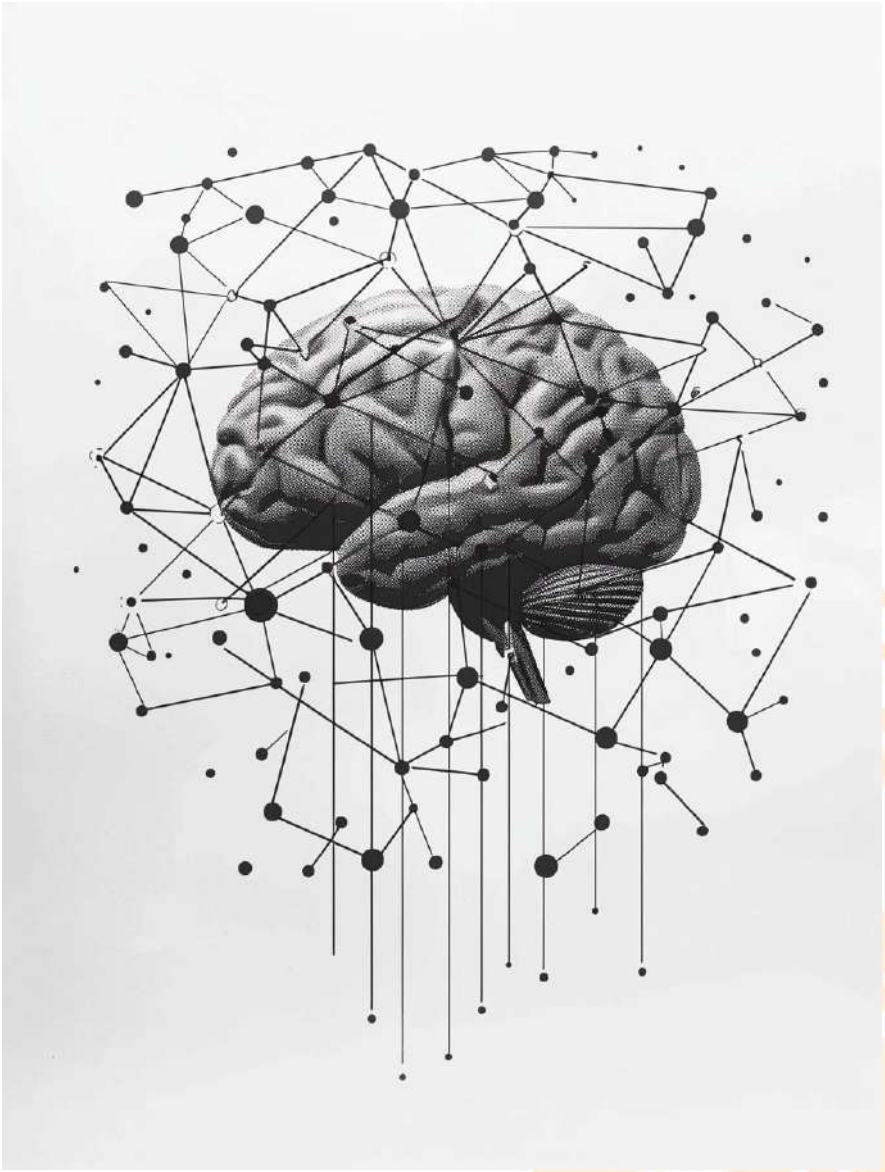
Sí, y se deben fundamentalmente a problemas por:

- **Generalización** OOD
- **Eficiencia muestral**
- Manejo **causalidad**

Ilustración creada con IA (Midjourney)

Dimensiones inteligencia humana

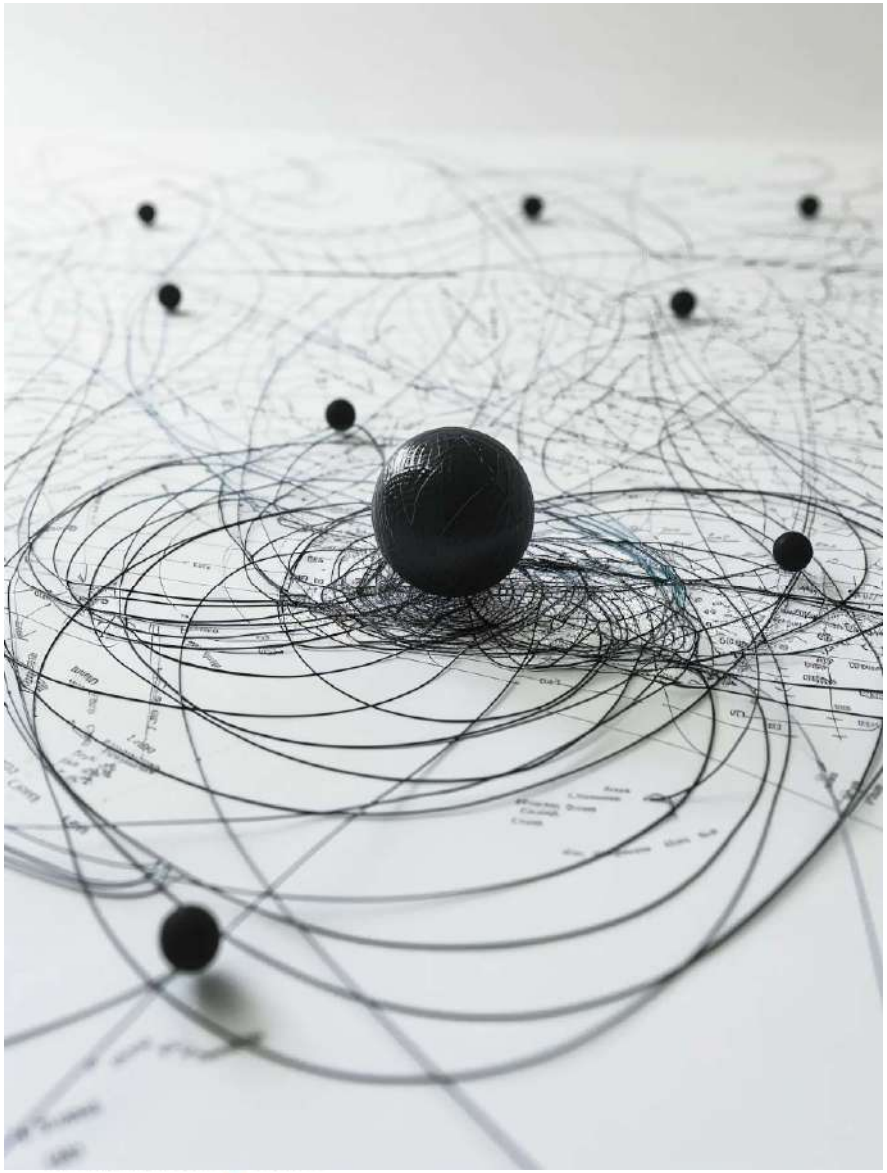
- Capacidad de crear y usar **modelos del mundo**
- **Memorización**
- **Razonamiento**
- **Planificación**



comillas.edu

Ilustración creada con IA (Midjourney)

Creación y uso de modelos del mundo



- Cierta **polémica** en la comunidad investigadora
- **Comportamientos** compatibles con ello
- Provienen del **lenguaje (imperfectos)**
- **Dificultad** para crearlos en imágenes **a nivel pixel**

Ilustración creada con IA (Midjourney)

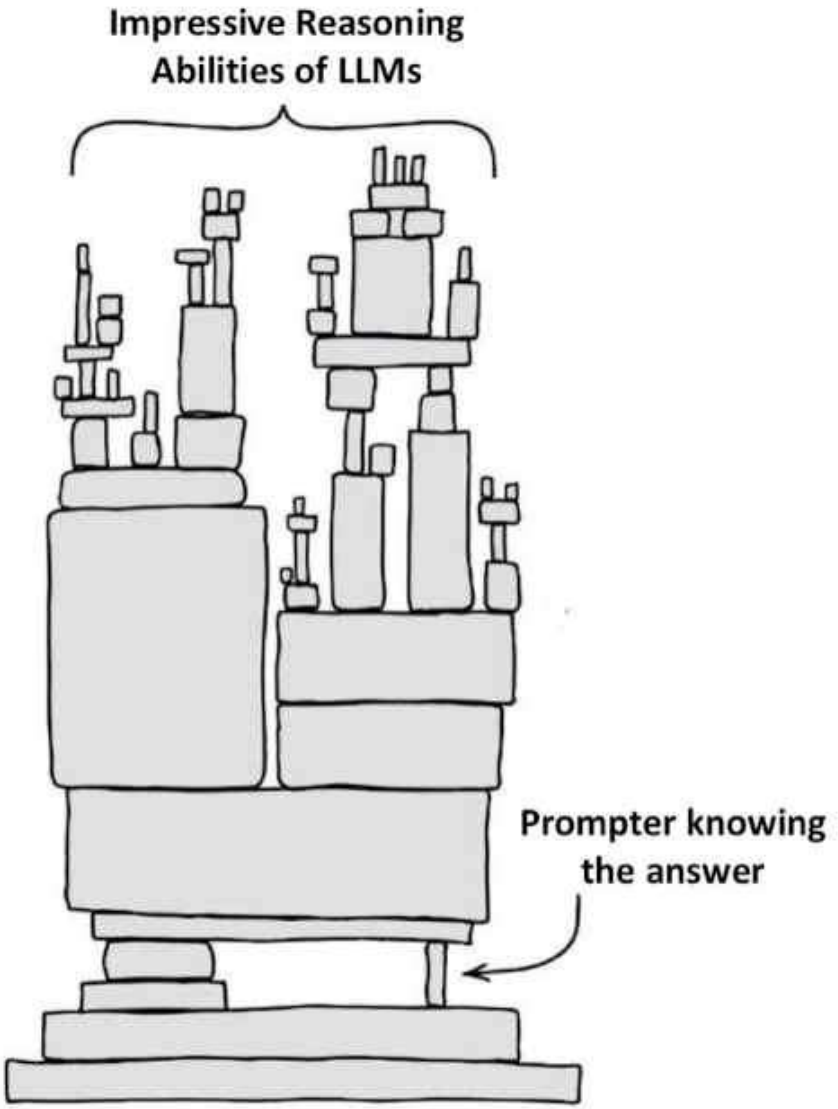
Memorización

- En esta dimensión **sí** hay un **buen funcionamiento**
- Arquitectura de **transformer → ventana de contexto**
- Además, **RL → memorización contextual**

comillas.edu

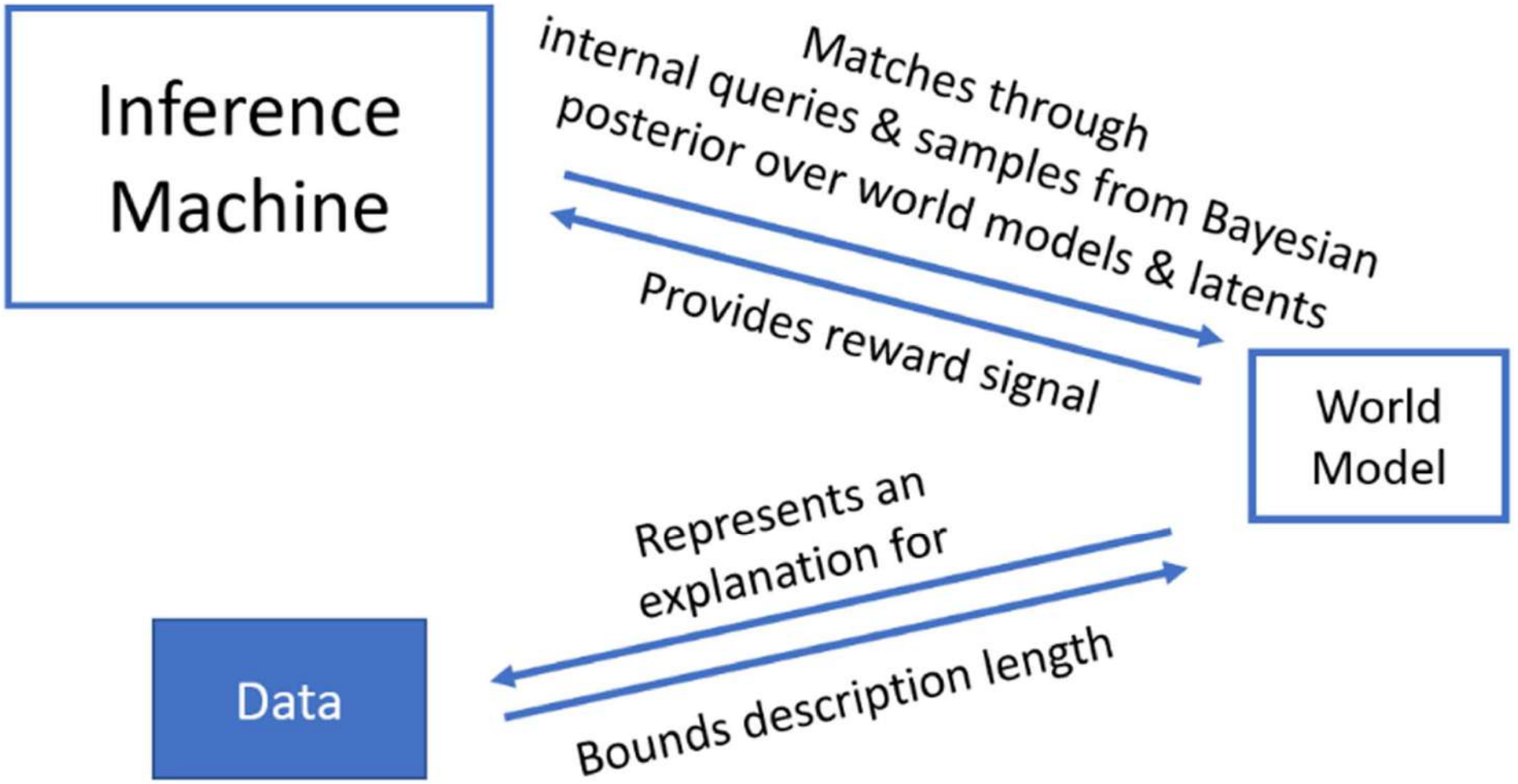
Ilustración creada con IA (Midjourney)

Razonamiento y planificación



Fuente: https://cacm.acm.org/blogcacm/can-llms-really-reason-and-plan/

Razonamiento y planificación



Fuente: https://yoshuabengio.org/2023/03/21/scaling-in-the-service-of-reasoning-model-based-ml/

¿Qué se está haciendo? Recordamos algunos retos

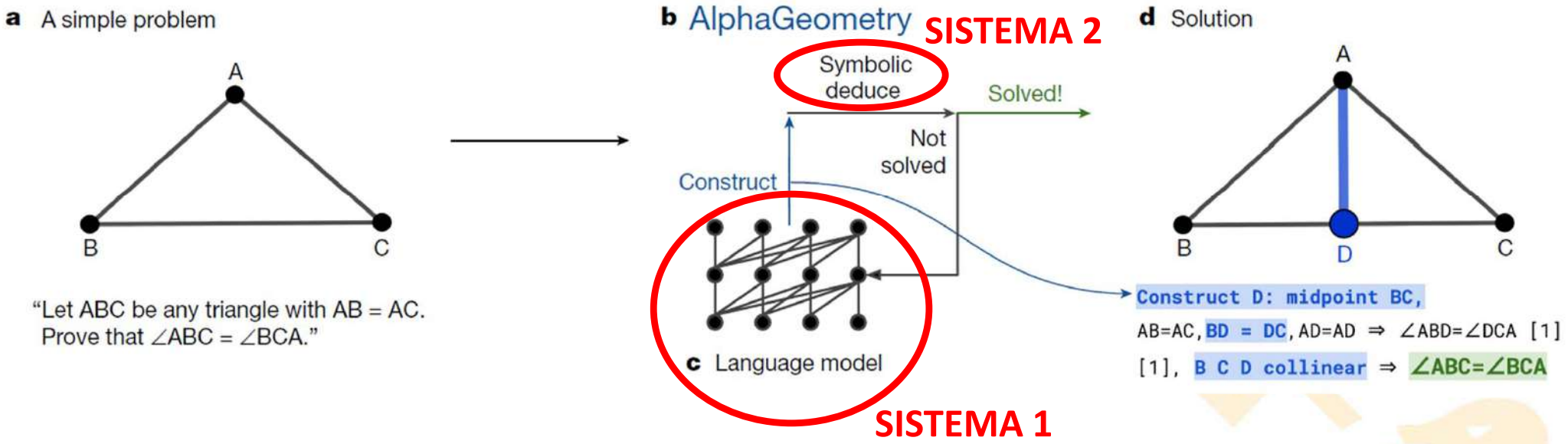
- Inconvenientes de operar en **espacios "en crudo"** (e.g. píxels)
  - Necesidad de desarrollar **abstracciones** eficaces

- **Escasez** de ejemplos
  - Muestras **sintéticas** ➔ Explotación en el sistema real

- **Difícultad** con **razonamiento y planificación**
  - **Sistema 2** de "pensar rápido, pensar despacio"

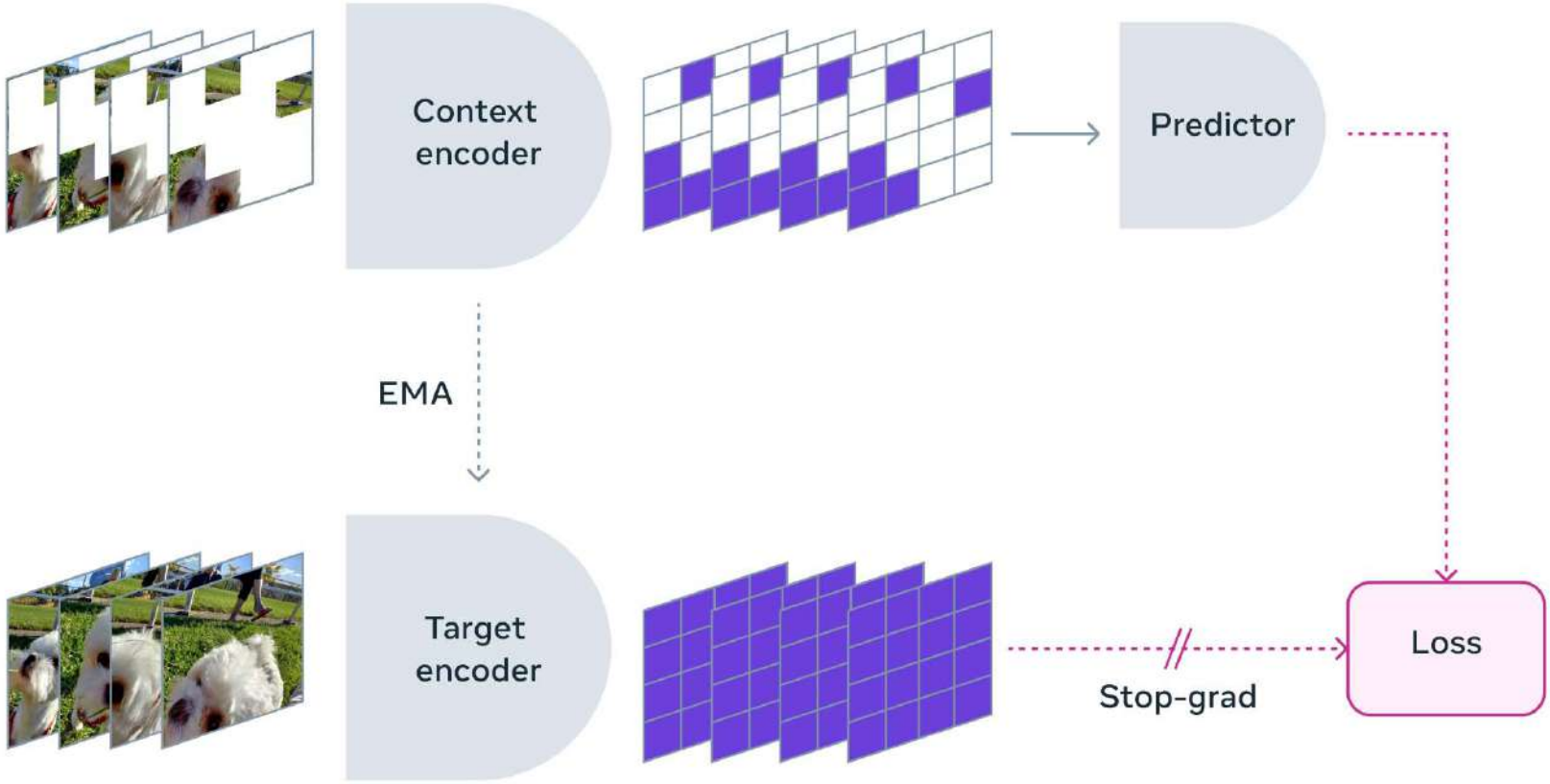
¿Qué se está haciendo? AlphaGeometry

- "Akin to the idea of "thinking, fast and slow", one system provides fast, "intuitive" ideas, and the other, more deliberate, rational decision-making."



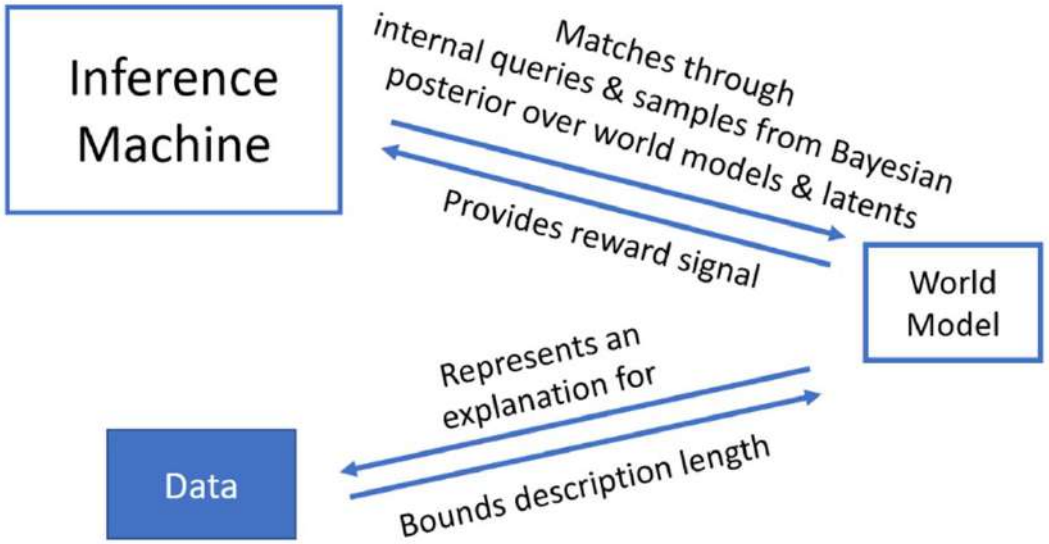
- **Aprende a partir de muestras sintéticas**, sin ningún ejemplo real
- Ideas similares en **planificación robusta**

¿Qué se está haciendo? V-JEPA



¿Qué se está haciendo? G-Flow Nets

- Máquina de **inferencia "amortizada"**
- Funcionamiento **composicional**
- Capaz de **generar preguntas** para mejorar modelo modular del mundo de forma eficiente.
- Aplicaciones:



  - Descubrimiento/**diseño de moléculas** ("secuencias biológicas")
  - **Planificación** robusta
  - Diseño asistido por IA de **experimentos** científicos