

**Semana 7 - Sesión 1**

# **Verificar antes de creer**

Seminario sobre Enseñanza de las Matemáticas I

**“Entender cómo  
‘piensa’ la  
Inteligencia Artificial  
es el primer paso para  
no dejarnos engañar”**

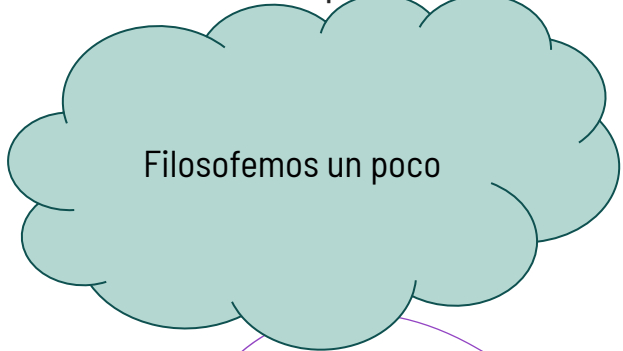
# **Información no es sinónimo de certeza**

# ¿Qué es LA VERDAD?

La IA no tiene  
criterios de verdad de  
fábrica

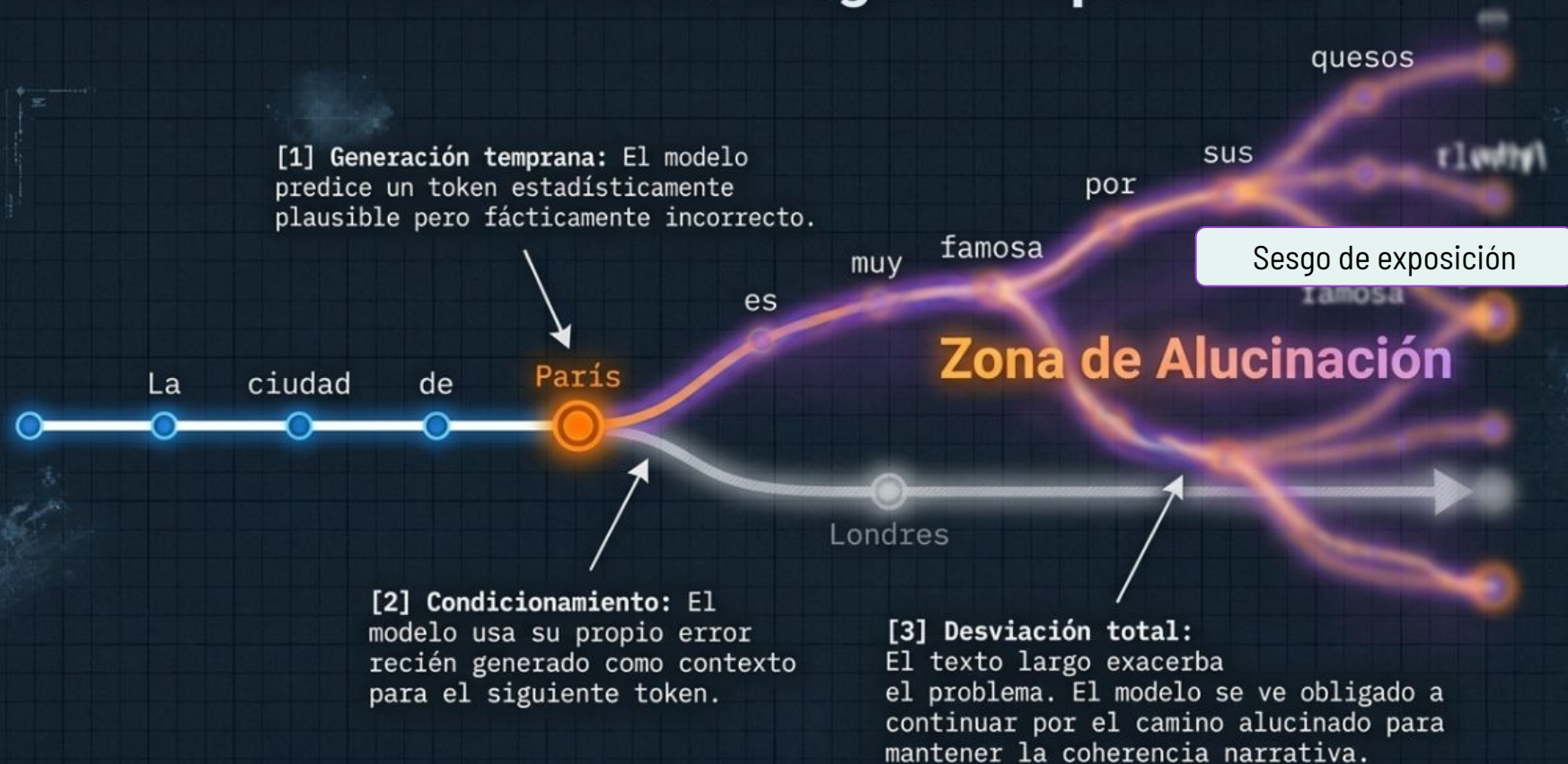
# VERDAD

- Aristóteles: la verdad ocurre cuando el juicio coincide con los hechos reales.
- Platón: idea perfecta y permanente que va más allá de lo que vemos.
- Corrientes escépticas: la verdad total es muy difícil o imposible de alcanzar por el ser humano.



Filosofemos un poco

# El efecto de bola de nieve: Sesgo de Exposición.



# La barrera de la "Cola Larga" (Long Tail).



Un modelo LLM, sin importar su tamaño, siempre luchará con hechos del Torso y la Cola. Escalar un modelo solo engrosa el conocimiento de la Cabeza, dejando la Cola expuesta al espejismo estadístico.

# Dos enfoques

## Desde el software

Chain-of-Verification /  
CoVe - Meta AI

## Desde el usuario

LOVE - Marco de  
alfabetización crítica

# CoVe

## Primera respuesta

Con todo y alucinaciones por defecto.

## Plan de verificación

"Rúbrica" de evaluación (sobre la primera respuesta).

## Variante factorizada

Aislar la consulta para evitar el sesgo de exposición / verificar el dato.

Evitar preguntas dicotómicas o sesgadas.

## Respuesta final

Compara respuesta vs borrador. Identifica contradicciones y genera respuesta unificada con datos verificados.

**“Si una súper computadora requiere un **proceso lógico** (y rígido) de 4 pasos aislados (y auditables) para **no mentirse a sí misma**, ¿qué esperaríamos para un estudiante?”**

Amigui, date cuenta



**L.O.V.E.**

Cortafuegos (firewall)  
cognitivo en las personas.

# Consecuencias de L.O.V.E.

- Lógica: Cuidar falacias lógicas generadas por información “correcta”, relacionadas de manera “incorrecta”.
- Originalidad: Preservar integridad intelectual. Asumir la responsabilidad de ser el autor, usando una herramienta o un “contribuyente no humano”. Cuidar los sesgos.
- Verificabilidad: Exige contrastar y validar rigurosamente los datos.
- Ética: Uso responsable considerando el impacto humano.