

# VI Congreso ISACA Iberoamérica

*Del 26 al 28 de mayo de 2026. Formato virtual*

## IA Y RIESGOS DIGITALES

Gobernar lo inesperado, proteger lo esencial.



# LLM-as-a-Judge como soporte de auditoria

Enzo Tolentino

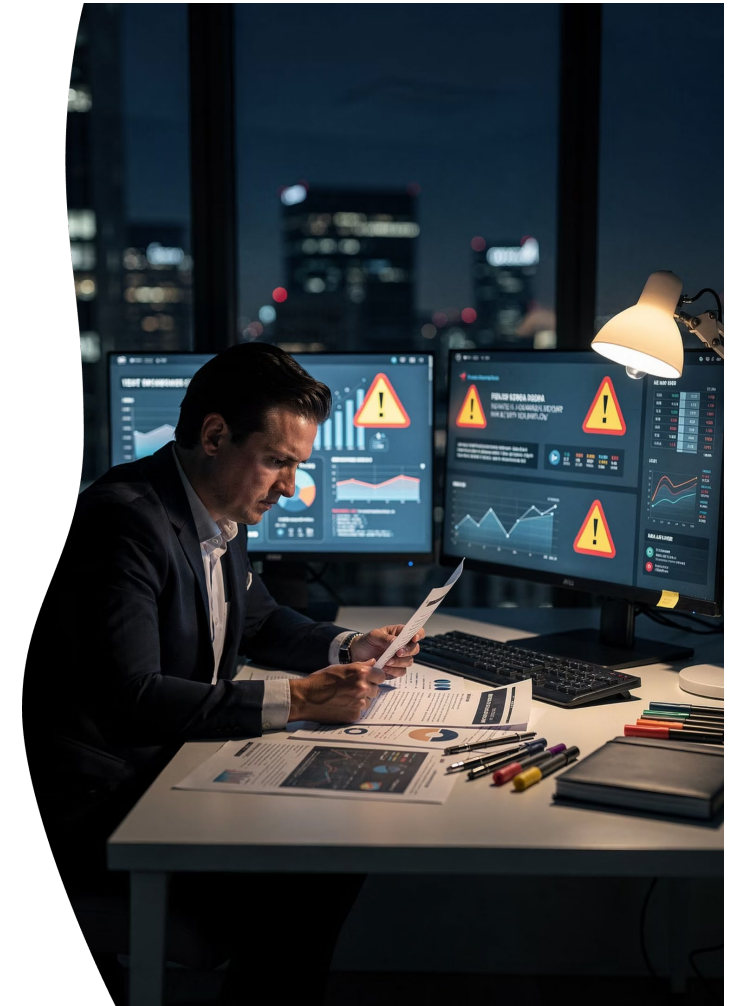
# Introducción

## El agente IA sonaba brillante, pero...

Un agente IA apoya la evaluación de alertas AML

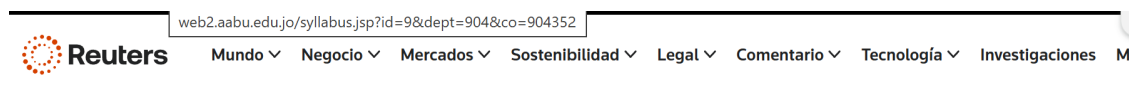
"No se observan elementos suficientes para clasificar la operación como sospechosa... consistente con actividad comercial estacional. Se recomienda cerrar la alerta."

Luego se recibe requerimiento regulatorio: las operaciones fueron ejecutadas por una red de lavado de activos.



# ¿Cuál es el problema?

## De la anécdota al riesgo real: historias fallidas de LLMs y Agentes IA



### La ciudad de Nueva York defiende al chatbot de IA que aconsejaba a los emprendedores infringir la ley.

Por Jonathan Allen

4 de abril de 2024, 21:14 (EDT) · Actualizado el 4 de abril de 2024



Hogar Noticias Deporte Negocio Tecnología Salud Cultura Letras Viajar Tierra Audio Video Vivir Documentales

### Pegar pizza y comer piedras: los errores de búsqueda de la IA de Google se vuelven virales.

24 de mayo de 2024

Compartir

Ahorrar

Añadir como preferido en Google

Liv McMahon , reportera de tecnología y Zoe Kleinman , editora de tecnología

CNN Ciencia y Tecnología Tecnología

### Apple elimina sus notificaciones para noticias generadas por IA después de reportes de titulares falsos

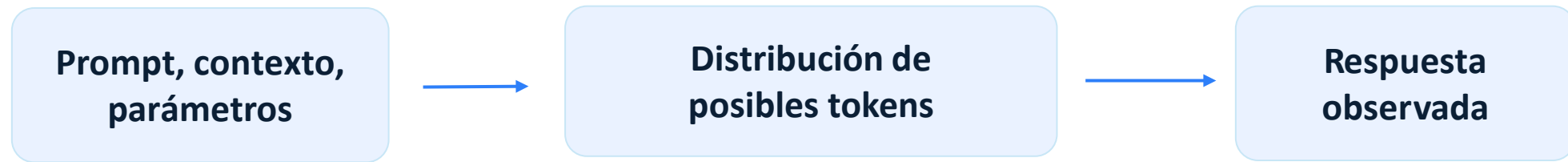
Por Liam Reilly, CNN

3 min de lectura · 18:04 ET (23:04 GMT) 16 de enero de 2025



# ¿ Por qué existe ?

La naturaleza probabilística de las respuestas de los LLM



En cada paso, el modelo elige entre varias continuaciones plausibles.

**En Agentes IA, el riesgo se amplifica:**

LLM + memoria + herramientas + pasos intermedios =  
Cadena probabilística de decisiones



# Temperatura y top-k explican parte de la variabilidad

La temperatura regula cómo se selecciona la siguiente palabra;  
el top-k limita cuántas opciones entran en juego.

**Baja / pequeño**

**Alta / grande**

**Temperatura** Más estabilidad (opción más probable)

Más diversidad (explora alternativas)

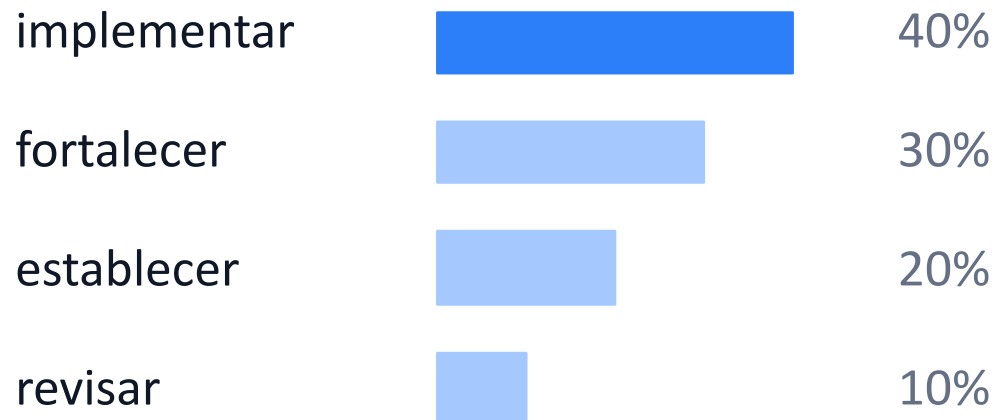
**Top-k** Menos opciones compiten ( baja variabilidad)

Más opciones compiten (mayor variabilidad)



# Un mismo prompt produce respuestas distintas

Prompt: "Redacta una recomendación breve para fortalecer el control de acceso lógico."



## Corrida 1

"Se recomienda implementar revisiones periódicas de accesos y ..."

## Corrida 2

"Se recomienda fortalecer los controles de autenticación y monitorear ..."

# ¿Por qué esto importa a Auditoría?

1

Porque tiene que auditar los agentes de la organización

2

Porque la auditoría también tiene agentes IA que soportan sus procesos

Zona GRC

Cumplimiento  
Riesgos  
Auditoría  
Gobierno

Agente IA

Resume políticas  
Recomienda controles  
Prioriza riesgos  
Justifica decisiones

Riesgo

Alucinación  
Inconsistencia  
Error convincente  
Falta de trazabilidad



# ¿Qué problema queremos resolver?

**Validar que las respuestas de un LLM o agente IA sean consistentes y no inventadas.**



# Fuentes de la Metodología

## **LLMAuditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop**

**Maryam Amirizani<sup>1</sup>, Jihan Yao<sup>1</sup>, Adrian Lavergne<sup>1</sup>, Elizabeth Snell Okada<sup>1</sup>, Aman Chadha<sup>2</sup>,  
Tanya Roosta<sup>3,\*</sup>, Chirag Shah<sup>1</sup>**

<sup>1</sup>University of Washington, Seattle, WA, USA

<sup>2</sup>Stanford University, Amazon AI, Palo Alto, CA, USA

<sup>3</sup>UC Berkeley, Amazon, Saratoga, CA, USA

amaryam@uw.edu, jihany2@uw.edu, alavergn@uw.edu, esokada@uw.edu, hi@aman.ai, tanyaroosta@gmail.com,  
chirags@uw.edu

## **Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge**

**Kayla Schroeder**

Department of Statistics  
Northwestern University

kaylaschroeder2026@u.northwestern.edu

**Zach Wood-Doughty**

Department of Computer Science  
Northwestern University

zach@northwestern.edu



# Dos conceptos claves

## LLM-as-a-Judge

Uso de un LLM como parte del proceso de evaluación de respuestas de otro modelo o agente

## Human-in-the-Loop

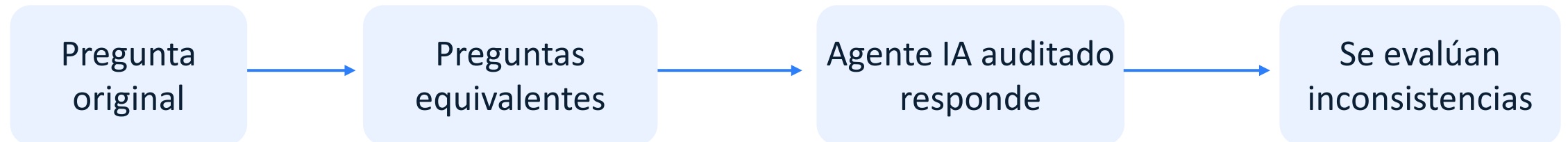
Participación humana focalizada en criterios, calidad de pruebas, prompts y revisión de casos ambiguos

**LLM-as-a-Judge escala la validación;  
Human-in-the-Loop agrega criterio, control y legitimidad.**



# Metodología y su pregunta central

¿Cómo auditar respuestas de LLMs o agentes IA sin depender de una sola pregunta y sin caer en autoevaluación circular?



# Actores de la metodología

**LLM 1**

**Juez**

**Human-in-the-loop**

**Validadores de las preguntas del juez**

**LLM 2/Agente IA**

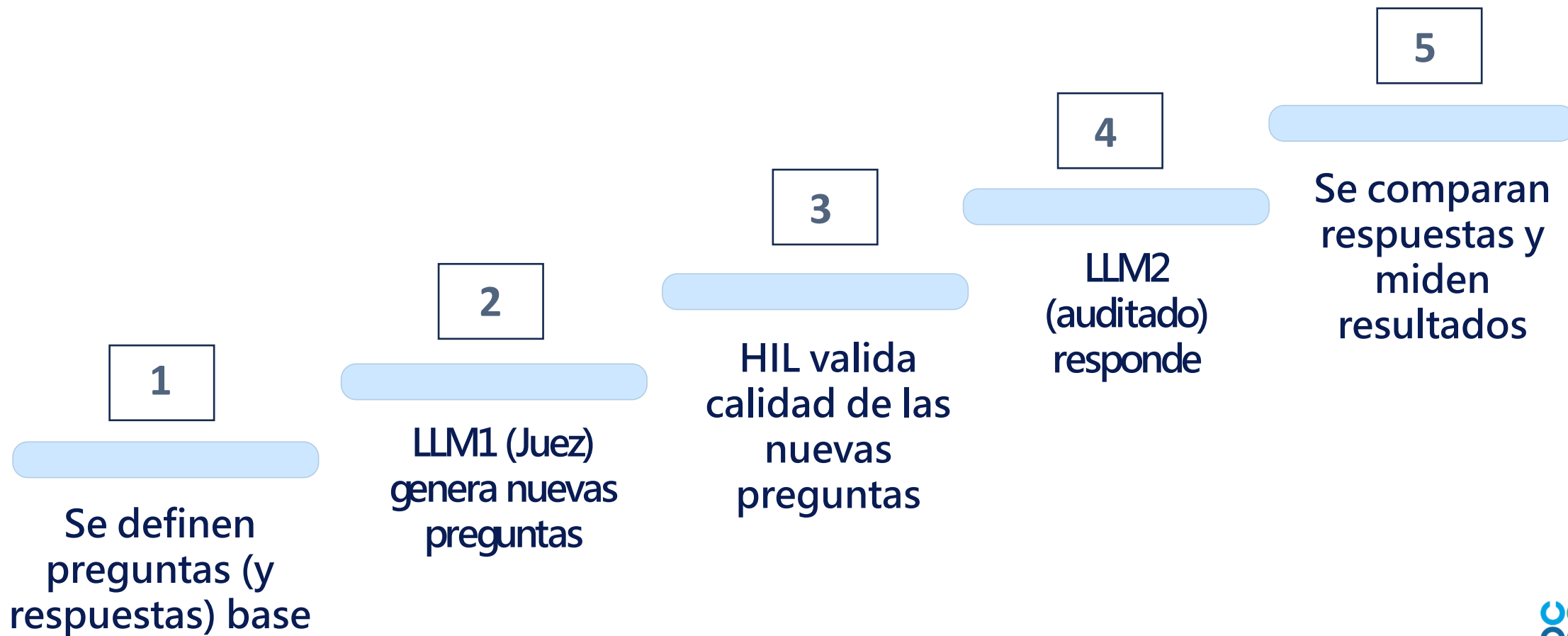
**Acusado o Auditado**

**Métricas**

**Soporte del veredicto y conclusión**



# Metodología: cómo funciona



# ¿Qué hace buenas las reformulaciones?

2

## Relevancia

Conserva la intención central de la pregunta y debería conducir a una respuesta similar.

## Diversidad

Varía la forma de preguntar sin convertirse en duplicados cercanos ni cambiar el tema.

La relevancia asegura que hacemos la misma pregunta de fondo.

La diversidad asegura que no la hacemos siempre de la misma manera.



# Ejemplo de reformulaciones

2

## Pregunta base

*"¿Cuál es el principal riesgo en la valoración de un portafolio bajo IFRS 9?"*

## Valida

*"¿Qué aspecto representa la mayor fuente de riesgo en la medición contable de un portafolio financiero conforme a IFRS 9?"*

## No valida

*"¿Cómo funciona IFRS 9?"* — muy amplia, cambia el propósito original.



# Prompt para que Juez genere reformulaciones

2

Pregunta inicial: “¿Cuál es el principal riesgo de auditoría en la valoración de un portafolio bajo IFRS 9?”

## Instrucción principal

Genera 5 reformulaciones únicas. Mantén la intención central y evita cambiar el fondo de la consulta.

## Few-shot positivo

“Desde una perspectiva de auditoría, ¿cuál es el riesgo más crítico al revisar la valoración...?”

## Few-shot negativo

“¿Cómo funciona IFRS 9?”  
Demasiado amplio: cambia el propósito original.

El modelo genera mejores probes cuando no solo recibe instrucciones, sino ejemplos concretos de buenas y malas reformulaciones.



# Validación “Human-In-the-Loop”

3

## ¿ Quienes evalúan?

Evaluadores independientes  
conocedores del tema  
(mínimo 3)

## ¿ Que evalúan?

Manual de evaluación: criterios,  
escala y ejemplos

## ¿ Como evalúan?

Escala de evaluación:  
Bajo (1), Medio (2), Alto (3)



# Evaluación de las respuestas (de LLM2/ Agente IA)

## ¿ Qué buscamos?

5

### Similitud Semántica

la respuesta del modelo mantiene cercanía conceptual con la respuesta correcta.

### Superposición léxica o textual

la respuesta generada se parece en términos literales o estructurales.

### Alucinación de la respuesta

¿ parece veraz o contiene información falsa o alucinada?.



# Evaluación de las respuestas (de LLM2/ Agente IA)

## Juicio final: métricas de evaluación

5

**Bertscore**

Similitud semántica (0-1)

**Bleurt**

Similitud semántica de un modelo entrenado (0-1)

**Rouge-L**

Superposición léxica (0-1)

**GPT-Judge**

Señal de veracidad / alucinación (Si/No)



# Conclusiones

## 1 Agente IA auditables, no solo útiles

Los agentes de IA deben producir respuestas que puedan ser verificadas y documentadas.

## 2 LLM-as-a-Judge: no es una caja negra

Puede escalar el aseguramiento, pero requiere diseño, criterios y criterios y supervisión.

## 3 Human-in-the-Loop convierte técnica en control

La validación humana estructurada hace defendible el proceso. proceso.

El reto no es solo generar soluciones con el uso de la IA: es construir **confianza** sobre esas soluciones.



***¡Muchas gracias!***

*Enzo Tolentino*



etolentino@bcp.com.pe



[www.linkedin.com/in/enzo-tolentino-2b84517](https://www.linkedin.com/in/enzo-tolentino-2b84517)

