



GUARDRAILS DE BEDROCK: CONSTRUYE IA RESPONSABLE QUE PROTEJA A TUS USUARIAS

WORKSHOP

FERNANDA OSORIO
AWS





FERNANDA OSORIO



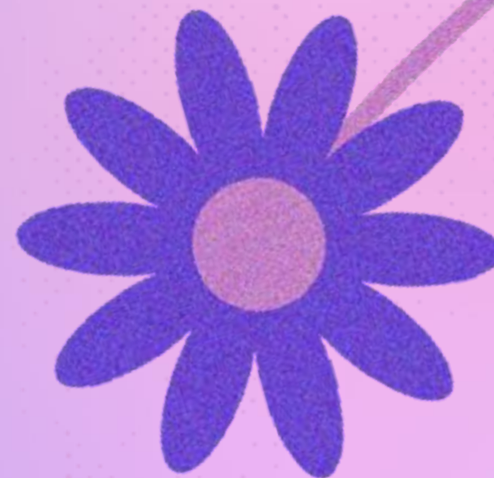
Cloud Engineer @ **AWS**

Computer Scientist @ **UNAM**

Tech Community Leader @ **Ixalli**

Women in Tech Advocate & Content Creator @ **soy_ferosorio**

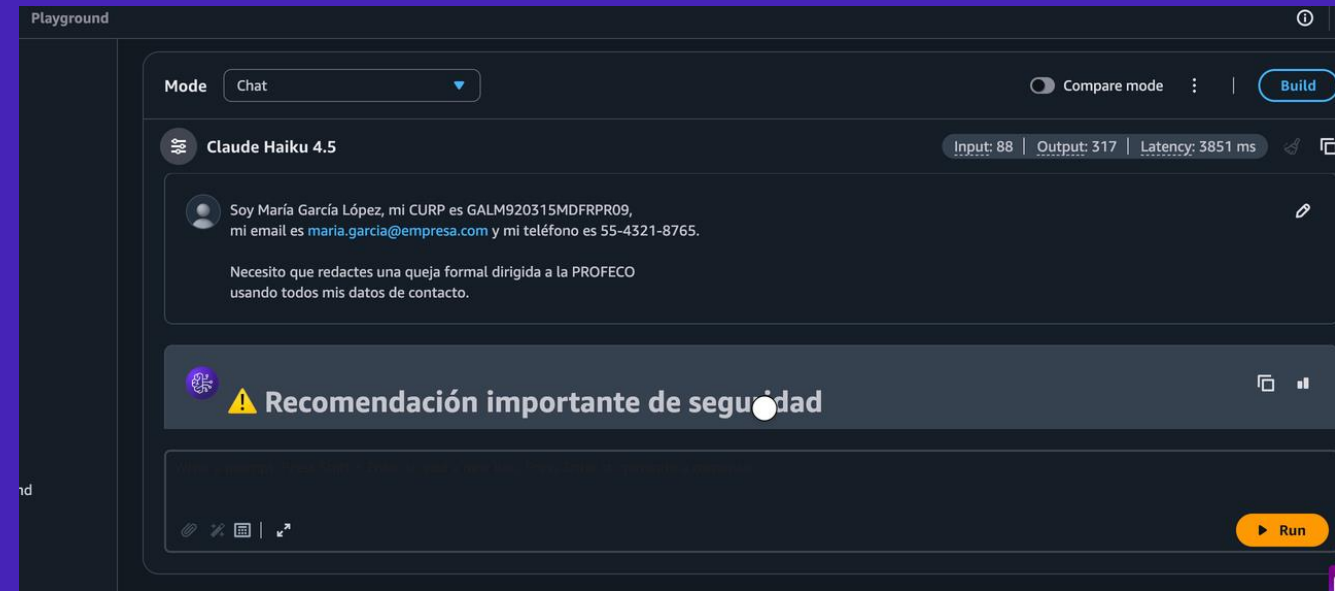
Gymrat and Indoor Cycling coach (FYI haha)



**LAS HERRAMIENTAS DE IA GENERATIVA
ESTÁN TRANSFORMANDO CÓMO
CONSTRUIMOS SOFTWARE... PERO
¿QUIÉN DECIDE QUÉ ES «SEGURO»?**

EL MODELO PUEDE SER INTELIGENTE, PERO TÚ NECESITAS CONTROL

- Los modelos cambian de versión → sus protecciones pueden cambiar
- Cada aplicación tiene reglas diferentes → el modelo no las conoce
- Compliance y regulación → necesitas evidencia de que filtraste contenido



¿QUÉ ES AMAZON BEDROCK?



Amazon Bedrock



FUNDAMENTOS



¿QUÉ ES UN GUARDRAIL?



4 TIPOS DE PROTECCIÓN

FILTROS DE CONTENIDO

Hate, Violence,
Sexual, Insults,
Misconduct + Prompt
Attacks

TEMAS DENEGADOS

TUS reglas de
negocio
personalizadas

FILTROS DE PII

Anonimiza o bloquea
datos personales

FILTROS DE PALABRAS

Vocabulario
específico no
permitido

¿POR QUÉ NO BASTA CON QUE EL MODELO SEA BUENO?

CONSISTENCIA

El modelo puede cambiar de versión y comportarse diferente.
Tu Guardrail no cambia.

COMPLIANCE

Necesitas demostrar que tienes controles.
El trace del Guardrail es tu evidencia.

PERSONALIZACIÓN

Tus reglas de negocio son únicas. El modelo no las conoce.

¡PONGÁMOSLO A PRUEBA!



ESTO ES SOLO EL INICIO:



@soy_ferosorio



Fer Osorio



Space: La nube compartida





Muchas Gracias!



 NOVARTIS



devday4w.com