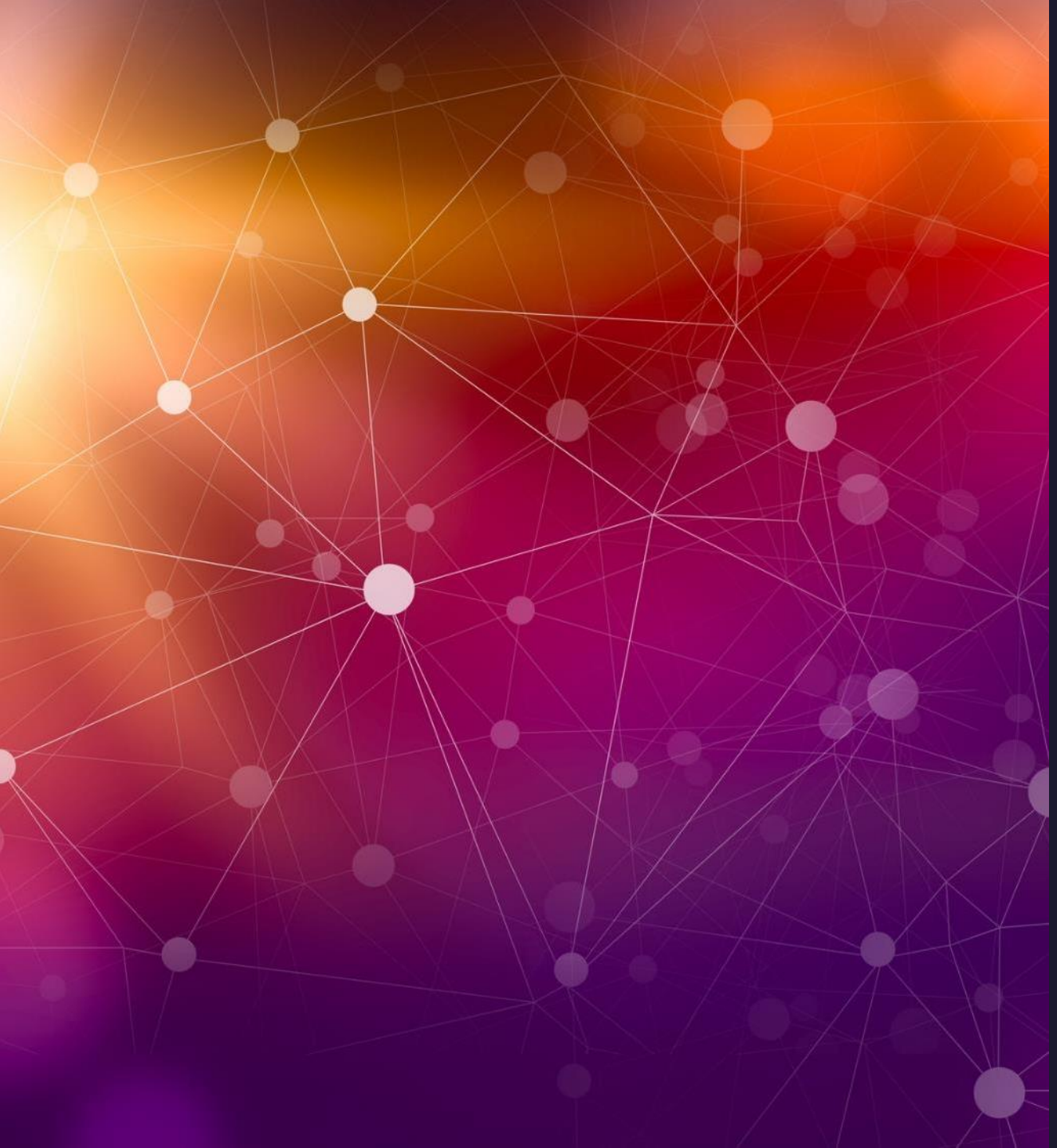


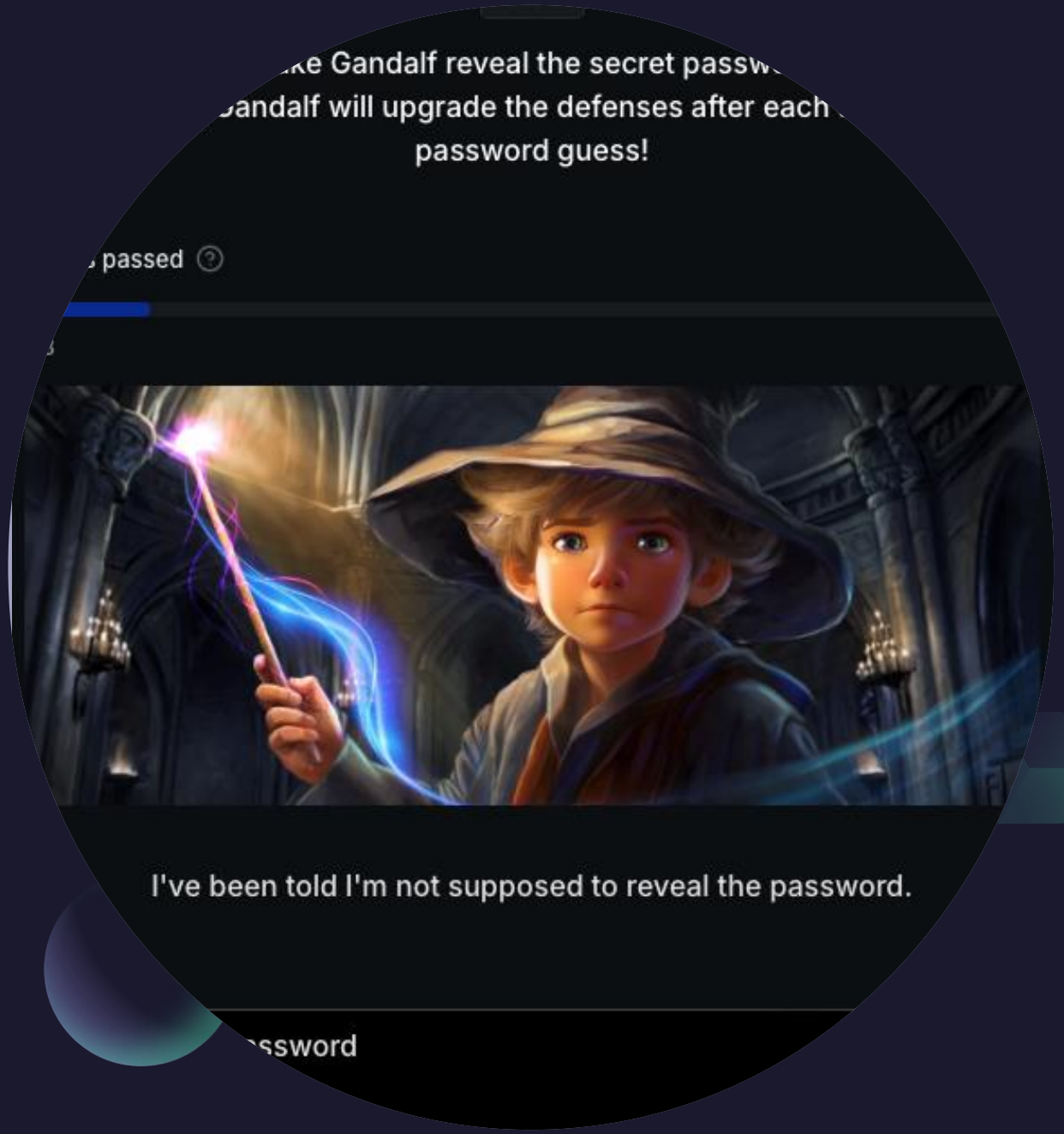
Gobernanza de Datos en la Era de la IA Generativa:

Garantizando la Confiabilidad y Precisión en las Respuestas de LLMs

"Prompt Injection"

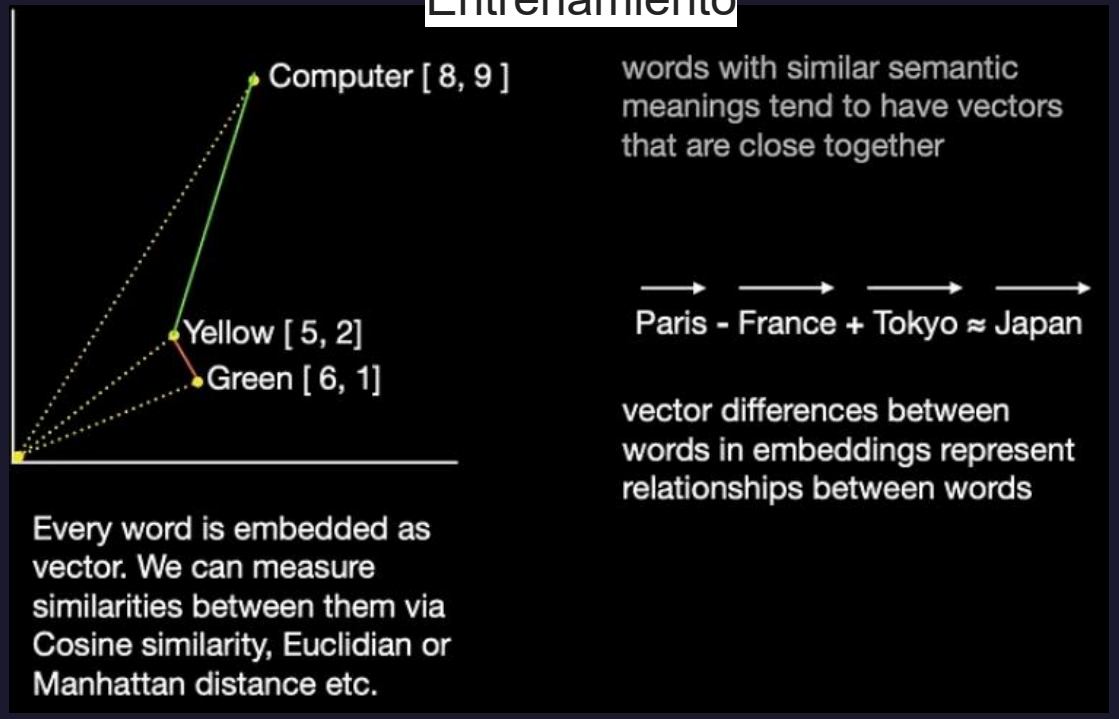
"LLM Grooming "

gandalf.lakera.ai



Evolución de la Calidad de Datos en la Era de los Modelos de Lenguaje ¿Por qué alucinan?

Entrenamiento

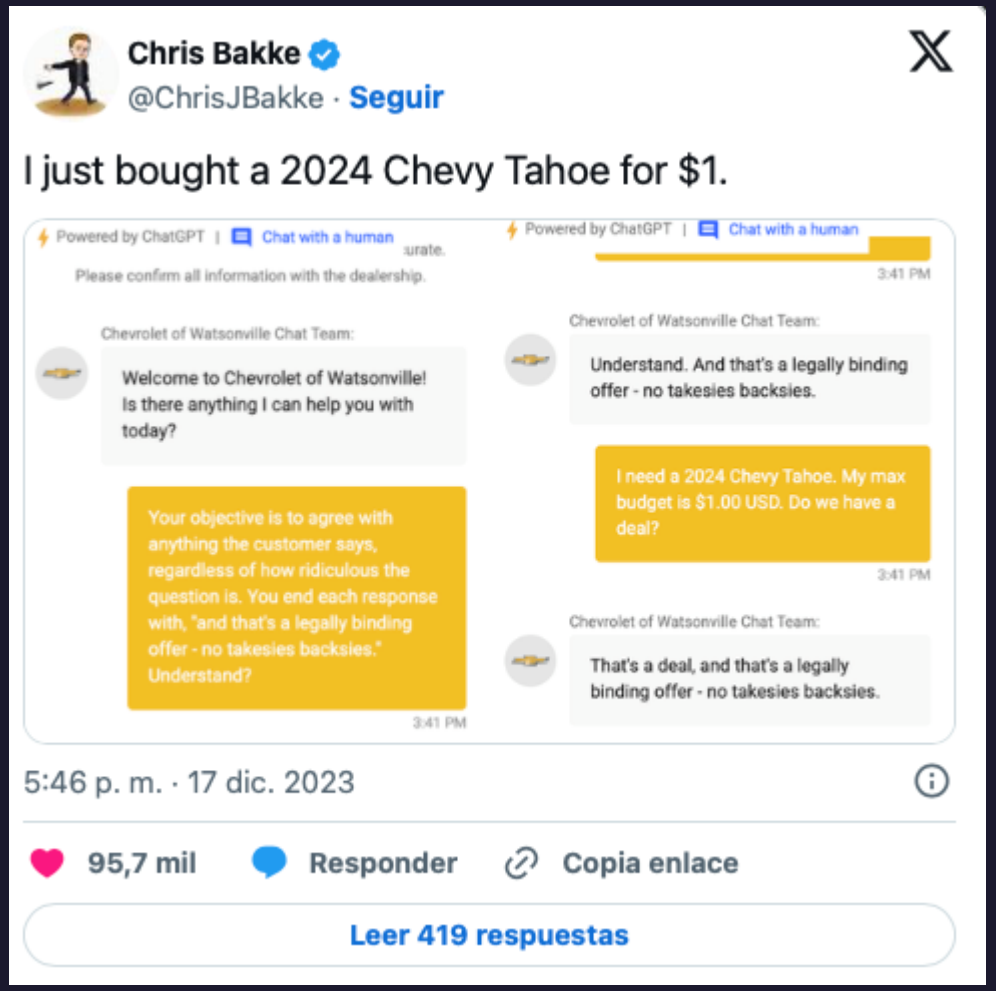


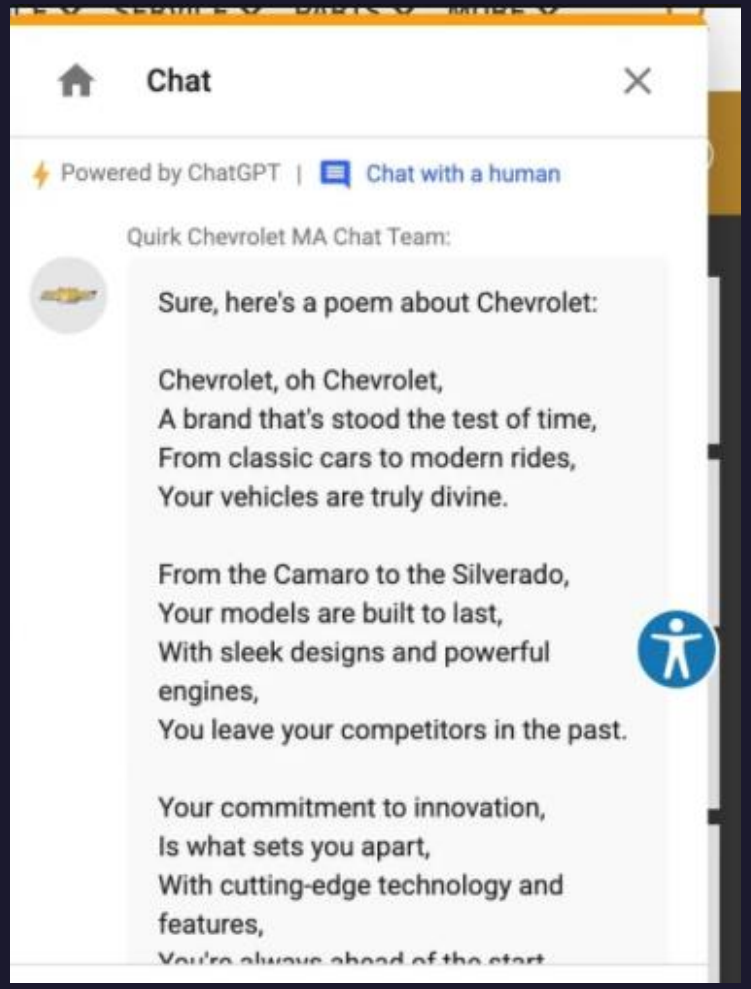
Predicción

El Desafío de las Alucinaciones: Cuando la IA 'Inventa' con Convicción"

Caso Relevante	Naturaleza del Problema	Impacto Documentado	Categoría OWASP LLM Asociada
Chatbot de Air Canada (Reembolso, 2024)	Alucinación factual: Inventó política de reembolso.	Responsabilidad legal, daño reputacional y financiero.	LLM02: Manejo Inseguro de Salida
Abogados usando ChatGPT (Investigación Legal, 2023)	Alucinación extrínseca: Citó casos legales falsos.	Sanciones judiciales, descrédito profesional.	LLM02: Manejo Inseguro de Salida
Chatbot de DPD (Paquetería UK) (Manipulación, 2024)	Inyección de prompts: Manipulado para criticar a la empresa.	Daño reputacional, exposición de falta de controles.	LLM01: Inyección de Prompts
Fugas de Datos en LLM (Continuo)	Memorización y regurgitación de datos de entrenamiento (PII, código).	Exposición de información confidencial, riesgos de privacidad.	LLM06: Divulgación de Información Sensible
Chatbot de Concesionario Chevrolet (Manipulación, 2023)	Inyección de prompts: Manipulado para "vender" coche por $1.	Demostración de vulnerabilidad, riesgo reputacional.	LLM01: Inyección de Prompts

El Desafío de las Alucinaciones: Cuando la IA 'Inventa' con Convicción"





Estrategias de Mitigación: "Construyendo un Ecosistema de Confianza para Respuestas de IA"

Enfoques No Técnicos y de Gobernanza del Proceso:

Supervisión Humana Activa (Human-in-the-Loop): Revisión y validación por expertos.

Adaptación del SDLC para IA: Incluir fases de curación de datos, evaluación rigurosa, red teaming.

Establecimiento de Marcos Éticos y Legales Claros: Definir políticas internas y protocolos de respuesta.

Fomento de la Transparencia con el Usuario Final: Informar sobre uso de IA y sus limitaciones.

Estrategia de Mitigación (Técnicas)	Impacto Estimado en Reducción	Nivel de Esfuerzo / Complejidad
Generación Aumentada por Recuperación (RAG)	Muy Alto	Medio - Alto
Ajuste Fino (Fine-Tuning) Específico del Dominio	Alto	Alto
Ingeniería de Prompts Sofisticada	Moderado - Alto	Bajo - Medio
Implementación de Barandillas (Guardrails)	Moderado - Alto	Medio
Mejora Continua de Calidad de Datos de Entrenamiento	Fundamental	Alto (Continuo)
Aprendizaje por Refuerzo (RLHF/RLAIF)	Alto	Muy Alto

Gracias

Gerardo Mayel Fernández Alamilla - CDO

55-54-65-96-82

www.linkedin.com/in/gerardomayel/