Aplicaciones LLM

Soluciones de LLM Ops

Soluciones de LLMOps en el mercado

- El mercado de las herramientas de LLMOps es muy reciente y está en plena ebullición, con nuevas alternativas apareciendo cada poco tiempo.

- Hay herramientas integrales más integrales como WhyLabs y otras más específicas como Guardrails AI. En esta lección analizaremos dos productos populares de WhyLabs: LangKit y LLM Security Management.

- Mientras que LangKit se centra en la extracción de insights accionables para la moderación de contenido y la observabilidad, LLM Security Management se enfoca en proteger las aplicaciones LLM contra una gama más amplia de riesgos de seguridad, incluyendo inyecciones de prompts, fugas de datos y desinformación.

Features principales de LangKit, de WhyLabs

- Utiliza técnicas de lenguaje natural para extraer insights accionables de los prompts y respuestas, identificando y mitigando prompts maliciosos, datos sensibles, respuestas tóxicas, temas problemáticos, alucinaciones e intentos de jailbreak.
- Permite definir límites y detectar prompts y respuestas problemáticos en tiempo real, tomando acciones apropiadas en caso de fallos.
- Valida cómo los LLM responden a prompts conocidos, tanto de manera continua como ad-hoc, para asegurar consistencia al modificar prompts o cambiar modelos.
- Extrae datos telemétricos clave y los compara con baselines inteligentes a lo largo del tiempo, ayudando en la depuración y el ajuste fino de la aplicación LLM.
- Se integra fácilmente con APIs públicas o modelos propietarios.
- Proporciona más de 50 señales de telemetría para evaluar la calidad, relevancia, sentimiento y seguridad de los prompts y respuestas.

Features de LLM Security Management, de Whylabs

- Protección contra Ataques Maliciosos:.
- Prevención de Fugas de Datos.
- Defensa contra Inyecciones de Prompts.
- Mitigación de la Desinformación.
- Adopción de Mejores Prácticas de Seguridad: Implementa la telemetría para capturar los riesgos de seguridad definidos en el "OWASP Top 10 para Aplicaciones LLM", permitiendo guardrails en línea, evaluaciones continuas y observabilidad.
- Manejo de Diversos Riesgos de Seguridad: Aborda una gama de vulnerabilidades, incluyendo manejo inseguro de salidas, envenenamiento de datos de entrenamiento, denegación de servicio, problemas en la cadena de suministro y excesiva dependencia en los LLMs.
- Guardrails y Registro Personalizable: Implementa guardrails en línea con métricas, umbrales y acciones personalizables, y registra cada par de prompt/respuesta.