Suplemento Técnico Avanzado: Dominio 4

IA Responsable en la Era Generativa, RLHF y Gobernanza

Material de Preparación para AWS AI Practitioner

1. Desafíos Únicos de la IA Generativa

A diferencia del ML tradicional (que resuelve problemas estrechos como el riesgo crediticio), la IA generativa es de "propósito general", lo que dificulta la definición y medición de la equidad y la seguridad.

1.1. Riesgos Emergentes

- **Toxicidad Subjetiva:** El contenido ofensivo puede ser sutil o depender del contexto cultural, lo que dificulta su detección automática.
- **Alucinaciones Fácticas:** Los LLMs generan secuencias probables de palabras, no necesariamente verdaderas. Ejemplo: creación de citas científicas o noticias financieras inexistentes que parecen veraces.
- **Propiedad Intelectual (IP):** El mimetismo de estilos artísticos o literarios y la regurgitación literal de fragmentos de código del dataset de entrenamiento plantean dilemas legales.
- **Plagio y Engaño:** Dificultad para verificar si un contenido fue creado por un humano o una IA, afectando sectores como la educación y los recursos humanos.

2. Técnicas de Mitigación Avanzadas

2.1. Protección de Datos y Modelos

- **Model Disgorgement (Vaciado de Modelo):** Proceso para eliminar los efectos de datos protegidos o específicos del resultado del modelo sin reentrenarlo por completo.
- **Privacidad Diferencial:** Técnica de entrenamiento que asegura que la presencia o ausencia de un dato individual en el entrenamiento no afecte significativamente la salida, evitando filtraciones de PII.
- **Marcas de Agua (Watermarking):** El modelo divide internamente los posibles tokens en listas "verdesz rojas". Al elegir solo de la lista verde, se crea una prueba estadística de que el texto fue generado por IA, indetectable para el humano.

3. RLHF: Alineación con Valores Humanos

El Aprendizaje por Refuerzo a partir de la Retroalimentación Humana (RLHF) es la técnica estándar para que la IA sea **amable, honesta e inocente**.

3.1. El Proceso de 4 Etapas

1. **Recopilación de Datos:** Humanos generan respuestas ideales para prompts específicos.
2. **Ajuste Fino Supervisado (SFT):** Se entrena un modelo base con el dataset anterior para que aprenda el formato de respuesta humana.
3. **Modelo de Recompensa (Reward Model):** Los humanos clasifican (rank) múltiples respuestas del modelo de mejor a peor. Se entrena una IA separada para predecir qué puntuación daría un humano a una respuesta.
4. **Optimización por RL:** El modelo principal se ajusta iterativamente usando el Modelo de Recompensa para maximizar su "puntuación humana".

4. Mejores Prácticas y Gobernanza de AWS

Para implementar una IA responsable, las organizaciones deben adoptar un enfoque centrado en las personas y en la arquitectura.

4.1. Estrategias Organizacionales

- **Equipos Multidisciplinarios:** Incluir no solo técnicos, sino especialistas en ética, legales y expertos en el dominio para identificar sesgos ciegos.
- **Niveles de Intervención Humana:**
  - **Human-in-the-loop:** El humano participa activamente en cada decisión.
  - **Human-on-the-loop:** El humano supervisa el sistema y puede intervenir si detecta errores.
  - **Human-over-the-loop:** El humano tiene la autoridad final sobre el sistema pero no interviene en cada proceso.
- **Uso de Casos Específicos:** AWS recomienda evitar aplicaciones çomodín"(catch-all) y definir casos de uso concretos para facilitar la trazabilidad y la rendición de cuentas.

4.2. Herramientas de Transparencia

- **AWS AI Service Cards:** Documentos que detallan cómo se usa el ML, sus limitaciones y consideraciones de diseño responsable para servicios específicos (ej. Rekognition, Transcribe).
- **SageMaker Model Cards:** Registro automático del "linaje"del modelo: cómo se entrenó, con qué datos y cuáles fueron sus métricas de evaluación.

5. Resumen de Valores Fundamentales

Objetivos de la Alineación

- **Amabilidad (Helpfulness):** El modelo debe ser útil para el usuario.
- **Honestidad (Honesty):** El modelo debe ser veraz y admitir sus limitaciones.
- **Inocencia (Harmlessness):** El modelo debe evitar generar contenido discriminatorio, ilegal o violento.