Manual de Estudio: AWS Certified AI Practitioner

Dominio 3.4: Evaluación del Rendimiento de Modelos Fundacionales

Material de Preparación Detallado

1. Desafíos del Despliegue y Optimización

Integrar un Modelo Fundacional (FM) requiere equilibrar el rendimiento técnico con las limitaciones de negocio.

1.1. El Equilibrio de la Inferencia

Al desplegar en la nube o en dispositivos periféricos (*edge*), se deben considerar tres factores:

- **Latencia:** Velocidad con la que el modelo genera la respuesta.
- **Computación y Almacenamiento:** Presupuesto de hardware disponible.
- **Rendimiento del Modelo:** Calidad de las respuestas.

1.2. Técnicas de Optimización

- **Reducción del tamaño del modelo:** Un modelo más pequeño carga más rápido y reduce la latencia, pero puede disminuir la exactitud.
- **Peticiones concisas:** prompts más cortos consumen menos ventana de contexto y procesan más rápido.
- **Ajuste de RAG:** Reducir el número y tamaño de los fragmentos (*chunks*) recuperados de la base de datos vectorial.

2. Métricas de Evaluación para IA Generativa

Dado que los resultados no son deterministas, no podemos usar métricas simples de acierto/error en todas las tareas.

2.1. Métricas Específicas de Tarea

3. Puntos de Referencia (Benchmarks) Globales

Para comparar modelos de propósito general, se utilizan conjuntos de datos estandarizados:

Métrica	Caso de Uso Principal
ROUGE	Evalúa resúmenes automáticos y traducción comparando el resultado generado con una referencia humana (*Recall-Oriented*).
BLEU	Algoritmo estándar para medir la calidad de la traducción automática entre idiomas naturales.
BERTscore	Calcula la similitud semántica utilizando modelos BERT; disponible en **Amazon Bedrock** para evaluar fidelidad y alucinaciones.

- **GLUE / SuperGlue:** Colección de tareas para evaluar la comprensión general del lenguaje (análisis de sentimiento, Q&A).
- **MMLU (Massive Multitask Language Understanding):** Evalúa conocimientos del mundo y resolución de problemas en temas como leyes, historia y matemáticas.
- **HELM (Holistic Evaluation of Language Models):** Ofrece transparencia y guía sobre qué modelo es mejor para cada tarea específica.
- **BIG-bench:** Tareas extremadamente complejas que desafían las capacidades actuales (biología, desarrollo de software, razonamiento lógico).

4. Herramientas de Evaluación en AWS

AWS facilita la medición mediante servicios administrados:

- **Amazon SageMaker Clarify:** Permite crear trabajos de evaluación para comparar métricas de calidad entre modelos de SageMaker JumpStart.
- **Evaluación de Bedrock:** Módulo automático para calcular puntuaciones de similitud semántica frente a referencias humanas.
- **Evaluación Humana:** Uso de trabajadores reales para comparar respuestas y determinar la preferencia humana.

5. Arquitectura de la Aplicación (La Pila de GenAI)

Para cumplir objetivos empresariales, la aplicación se organiza en capas integradas:

5.1. Las Capas de la Pila

1. **Capa de Infraestructura:** Proporciona computación, red y almacenamiento seguro (S3).
2. **Capa de Modelo:** Selección de los LLMs adecuados y configuración de inferencia.
3. **Capa de Herramientas y Orquestación:** Marcos de trabajo (como LangChain) para conectar el modelo con **RAG** o bases de datos externas.
4. **Capa de Aplicación:** Interfaz de usuario final (Sitio web o API REST).

6. Consideraciones de Almacenamiento y RAG

¿Por qué necesitamos almacenamiento adicional en aplicaciones de GenAI?

- **Recopilación de comentarios:** Guardar las finalizaciones de los usuarios para futuros procesos de **refinamiento** (*fine-tuning*) o alineación (*RLHF*).
- **RAG (Generación Aumentada por Recuperación):** Fundamental para evitar el çonocimiento desactualizado". Proporciona contexto externo en tiempo de inferencia, fundamentando las respuestas en datos reales y reduciendo alucinaciones.