Manual de Referencia Técnica: Dominio 3

Orquestación, Inferencia y Almacenamiento Vectorial Avanzado

Material Especializado para AWS AI Practitioner

1. Ecosistema de Modelos Fundacionales (FM)

Los Modelos Fundacionales son "generalistas". Para convertirlos en .^{es}pecialistas", AWS ofrece diversas familias con capacidades distintas.

1.1. Amazon Nova: La Nueva Generación

- **Modelos de Comprensión (Micro, Lite, Pro):** Aceptan texto, imagen y vídeo. Están optimizados para diferentes equilibrios entre velocidad y precisión.
- **Modelos de Generación Creativa (Canvas, Reel):** Diseñados específicamente para producir contenido visual (imágenes y vídeos) de alta calidad.

1.2. Modelos de Terceros en Bedrock

- **Claude 3.5 Sonnet (Anthropic):** Destaca en razonamiento complejo y supera a versiones anteriores en casi todos los benchmarks.
- **Stable Diffusion (Stability AI):** Especializado en texto-a-imagen utilizando un **espacio latente** para reducir el coste computacional.
- **Jurassic-2 (AI21 Labs):** Optimizado para seguir instrucciones complejas y generar texto con fluidez humana.

2. Control Matemático de la Inferencia

Al ejecutar inferencias, ajustamos parámetros que modifican la **función de masa de probabilidad** del siguiente token.

2.1. Aleatoriedad y Diversidad

- **Temperatura:** Modula la distribución. Una temperatura baja .^{a}gudiza"la curva (respuestas deterministas); una alta la .^{a}plana"(respuestas creativas).
- **Top K:** Filtra los K tokens más probables. Si $K = 50$, el modelo ignora el resto de la distribución.
- **Top P (Nucleus Sampling):** El modelo suma las probabilidades de los tokens más probables hasta llegar a P. Solo elige de ese subconjunto, permitiendo una diversidad dinámica.

2.2. Penalizaciones y Parada

- **Presence Penalty:** Evita que el modelo repita los mismos temas.
- **Frequency Penalty:** Reduce la probabilidad de repetir las mismas palabras exactas.
- **Stop Sequences:** Cadenas de texto que actúan como ïnterruptores"para que el modelo deje de generar contenido.

3. Arquitectura RAG y Bases de Conocimiento

La Generación Aumentada por Recuperación (RAG) resuelve el problema de las **alucinaciones** al proporcionar hechos externos al modelo.

3.1. El Proceso de Çhunking"

—
—
— —

Citas y Atribuci"80"363n Amazon Bedrock Knowledge Bases permite incluir citas en la respuesta generada. Esto es vital para aplicaciones empresariales donde el usuario debe poder verificar la fuente original de la informaci"80"363n.

Estrategias de B"80"372squeda Vectorial en AWS AWS est"80"341 integrando capacidades vectoriales en todos sus servicios de base de datos:

[h]
#I@*lp*8*cm*@*ServicioCapacidadVectorialAmazonOpenSearchSoportaalgoritmosHNSW*(*grafospara*
#I

4. Ingeniería de Peticiones (Técnicas de Razonamiento)

- **Chain of Thought (CoT):** Mejora el razonamiento lógico dividiendo problemas complejos en pasos intermedios.
- **Tree of Thoughts (ToT):** El modelo explora múltiples ramas de razonamiento y evalúa cuál es la más prometedora.
- **Least-to-Most:** El modelo primero desglosa una pregunta en sub-problemas más sencillos y los resuelve secuencialmente.
- **Mayéutica:** Se pide una explicación y luego se interroga esa explicación para descartar árboles de razonamiento incoherentes.

5. Métricas Críticas de Evaluación

- **Perplejidad (PPL):** Indica qué tan bien el modelo predice una muestra de prueba. **PPL bajo = Modelo más seguro y preciso**.
- **BERTscore:** Utiliza modelos de lenguaje para medir la **similitud semántica** entre la respuesta y la referencia, superando las limitaciones de BLEU que solo mira coincidencia de palabras.
- **Brevity Penalty (en BLEU):** Penaliza las traducciones que son demasiado cortas respecto a la referencia humana para evitar puntuaciones artificialmente altas.