

Tabla de contenido

Prefacio	xi
1. Casos de uso de IA generativa, fundamentos y ciclo de vida del proyecto	1
Casos de uso y tareas.....	1
Modelos básicos y centros de modelos.....	4
Ciclo de vida del proyecto de IA generativa.....	5
IA generativa en AWS.....	8
¿Por qué IA generativa en AWS?.....	11
Creación de aplicaciones de IA generativa con AWS.....	12
Resumen.....	13
2. Ingeniería de indicaciones y aprendizaje en contexto	15
Indicaciones y respuestas.....	15
Componentes léxicos.....	16
Ingeniería de indicaciones.....	16
Estructura de indicaciones.....	18
Instrucción.....	18
Contexto.....	18
Aprendizaje en contexto con inferencia con pocos golpes.....	20
Inferencia con cero golpes.....	21
Inferencia con un solo golpe.....	22
Inferencia con pocos golpes.....	22
Cuando el aprendizaje en contexto sale mal.....	23
Mejores prácticas de aprendizaje en contexto.....	24
Mejores prácticas de ingeniería de indicaciones.....	24
Parámetros de configuración de inferencia.....	29
Resumen.....	34
3. Modelos de lenguaje grandes básicos	35
Modelos de lenguaje grandes básicos.....	35
Analizadores léxicos.....	37
Vectores de incrustación.....	38
Transformadores.....	40
Ventana de entradas y contexto.....	41
Capa de incrustación.....	42
Codificador.....	42
Autoservicio.....	42
Decodificador.....	44

Salida softmax.....	44
Tipos de modelos básicos basados en transformadores.....	45
Conjuntos de datos para formación previa.....	48
Leyes de escalamiento.....	49
Modelos informáticos óptimos.....	51
Resumen.....	52
4. Optimizaciones de memoria y cálculo.....	55
Problemas de memoria.....	55
Tipos de datos y precisión numérica.....	58
Cuantificación.....	59
fp16.....	60
bfloat16.....	62
fp8.....	63
int8.....	64
Optimización de las capas de autoservicio.....	66
FlashAttention.....	66
Atención de consulta agrupada.....	67
Informática distribuida.....	68
Paralelismo de datos distribuidos.....	68
Paralelismo de datos totalmente fragmentados.....	70
Comparación del rendimiento de FSDP con DDP.....	72
Informática distribuida en AWS.....	74
Paralelismo de datos totalmente fragmentados con Amazon SageMaker.....	75
SDK de AWS Neuron y AWS Trainium.....	76
Resumen.....	77
5. Ajuste fino y evaluación.....	79
Ajuste fino de las instrucciones.....	80
Llama 2-Chat.....	80
Falcon-Chat.....	80
FLAN-T5.....	80
Conjunto de datos de instrucciones.....	81
Conjunto de datos de instrucciones multitarea.....	81
FLAN: ejemplo de conjunto de datos de instrucciones multitarea.....	82
Plantilla para indicaciones.....	83
Convierta un conjunto de datos personalizado en uno de instrucciones.....	84
Ajuste fino de las instrucciones.....	86
Amazon SageMaker Studio.....	87
Amazon SageMaker JumpStart.....	88
Amazon SageMaker Estimator para Hugging Face.....	89
Evaluación.....	90
Métricas de evaluación.....	91
Puntos de referencia y conjuntos de datos.....	92
Resumen.....	93

6. Ajuste fino con parámetros eficientes	95
Ajuste fino completo frente a PEFT	95
LoRA y QLoRA	98
Principios básicos de LoRA	99
Rango.....	100
Módulos y capas objetivo	100
Utilización de LoRA	101
Fusión del adaptador LoRA con el modelo original	103
Mantenimiento de adaptadores LoRA separados	103
Ajuste fino completo frente a rendimiento LoRA.....	104
QLoRA	105
Afinación de indicaciones e indicaciones suaves	106
Resumen.....	109
7. Ajuste fino con aprendizaje por refuerzo a partir	
de la retroalimentación humana	111
Alineación humana: útil, honesta e inofensiva.....	112
Panorama del aprendizaje por refuerzo.....	112
Forme un modelo de recompensa personalizado	115
Recopilación de datos de formación con personas en el ciclo.....	115
Ejemplo de instrucciones para etiquetadores humanos	115
Uso de Amazon SageMaker Ground Truth para anotaciones humanas	116
Prepare los datos de clasificación para formar un modelo de recompensa	118
Formar al modelo de recompensa	121
Modelo de recompensa existente: detector de toxicidad de Meta	122
Ajuste fino con aprendizaje por refuerzo a partir	
de la retroalimentación humana.....	124
Utilización del modelo de recompensa con RLHF.....	124
Algoritmo RL de optimización proximal de políticas	125
Ajuste fino del RLHF con PPO	126
Mitigación del pirateo de recompensas	128
Utilización del ajuste fino con parámetros eficientes con RLHF	130
Evaluar el modelo RLHF afinado	131
Evaluación cualitativa.....	131
Evaluación cuantitativa.....	131
Modelo de evaluación de la carga.....	132
Definición de la función de agregación métrica de evaluación	132
Comparación de las métricas de evaluación antes y después	133
Resumen.....	134
8. Optimización de la puesta en marcha de los modelos	137
Optimizaciones de modelos para inferencia.....	137
Poda	139
Cuantificación postformación con GPTQ	140
Destilación	142

Contenedor de inferencia de modelos grandes	144
AWS Inferentia: hardware específico para la inferencia	145
Estrategias de actualización e instalación de modelos.....	147
Pruebas A/B	148
Implementación en paralelo.....	149
Métricas y monitoreo	150
Autoescalado	151
Políticas de autoescalado	152
Definición de una política de autoescalado.....	152
Resumen.....	153
9. Aplicaciones de razonamiento sensibles al contexto usando RAG y agentes	155
Limitaciones de los modelos de lenguaje grandes.....	156
Alucinación.....	156
Corte de conocimiento.....	157
Generación mejorada por recuperación	157
Fuentes de conocimiento externas	158
Flujo de trabajo RAG.....	159
Carga de documentos.....	160
Agrupamiento	162
Recuperación y reordenación de documentos.....	163
Mejora de la indicación.....	164
Orquestación e implementación de RAG	165
Carga y agrupamiento de documentos	166
Almacenamiento y recuperación de vectores de incrustación	168
Cadenas de extracción.....	171
Reordenación con la relevancia marginal máxima	173
Agentes.....	175
Entorno ReAct.....	176
Entorno de lenguaje asistido por programas	178
Aplicaciones de IA generativa.....	182
FMOps: puesta en marcha del ciclo de vida del proyecto de IA generativa	188
Consideraciones sobre la experimentación.....	189
Consideraciones sobre el desarrollo.....	191
Consideraciones sobre la instalación en producción.....	192
Resumen.....	194
10. Modelos básicos multimodales	197
Casos de uso	198
Mejores prácticas de ingeniería de indicaciones multimodal.....	199
Generación y mejora de imágenes.....	200
Generación de imágenes.....	200
Edición y mejora de imágenes.....	201
Rellenado, pintura exterior, profundidad a la imagen.....	206
Rellenado.....	206

Pintura exterior	208
Profundidad a la imagen	209
Subtitulado de imagen y respuesta a preguntas visuales	211
Subtitulado de imagen.....	213
Moderación de contenido.....	213
Respuesta a preguntas visuales.....	213
Evaluación del modelo	218
Tareas generativas de texto a imagen.....	218
Difusión hacia delante.....	221
Razonamiento no verbal	221
Fundamentos de la arquitectura de difusión	223
Difusión hacia delante.....	223
Difusión hacia atrás	224
U-Net.....	225
Arquitectura de Stable Diffusion 2.....	226
Codificador de texto	227
U-Net y el proceso de difusión.....	228
Acondicionamiento de texto	230
Servicio combinado	230
Planificador.....	231
Decodificador de imagen.....	231
Arquitectura de Stable Diffusion XL.....	231
U-Net y el servicio combinado	232
Refinador.....	232
Acondicionamiento	233
Resumen.....	234
11. Generación controlada y ajuste fino con Stable Diffusion	237
ControlNet.....	237
Ajuste fino.....	242
DreamBooth	242
DreamBooth y PEFT-LoRA	245
Inversión textual	247
Alineación humana con aprendizaje por refuerzo a partir de la retroalimentación humana.....	251
Resumen.....	253
12. Amazon Bedrock: servicio gestionado para IA generativa	255
Modelos básicos de Bedrock	255
Modelos básicos de Amazon Titan.....	256
Modelos básicos de Stable Diffusion de Stability AI.....	256
API de inferencia de Bedrock.....	256
Modelos de lenguaje grandes	258
Generar código SQL	259
Resumir texto	259

Incrustaciones	260
Ajuste fino.....	263
Agentes.....	266
Modelos multimodales.....	269
Crear imágenes a partir de texto.....	269
Crear imágenes a partir de imágenes	271
Privacidad de datos y seguridad de la red.....	272
Gestión y monitoreo.....	273
Resumen.....	274