Engenharia de IA: construindo aplicações com modelos de fundação

SUMÁRIO

Prefácio

CAPÍTULO 1: Introdução à construção de aplicações de IA com modelos de fundação

- Surgimento da engenharia de IA
  - Dos modelos de linguagem aos grandes modelos de linguagem
  - Dos grandes modelos de linguagem aos modelos de fundação
  - Dos modelos de fundação à engenharia de IA
- Casos de uso dos modelos de fundação
  - Codificação
  - Produção de imagem e vídeo
  - Escrita
  - Educação
  - Bots de conversa
  - Agregação de informações
  - Organização de dados
  - Automação de fluxo de trabalho
- Planejando aplicações de IA
  - Avaliação do caso de uso
  - Definindo expectativas
  - Planejamento de etapas
  - Manutenção
- Pilha da engenharia de IA
  - As três camadas da pilha de IA
  - Engenharia de IA versus engenharia de ML
  - Engenharia de IA versus engenharia full-stack
- Resumo

CAPÍTULO 2: Entendendo os modelos de fundação

- Dados de treinamento
  - Modelos multilíngues
  - Modelos específicos de domínio
- Modelando
  - Arquitetura do modelo
  - Tamanho do modelo
- Pós-treinamento
  - Ajuste fino supervisionado
  - Ajuste fino conforme as preferências
- Amostragem
  - Aspectos básicos da amostragem
  - Estratégias de amostragem
  - Computação em tempo de teste
  - Saídas estruturadas

A natureza probabilística da IA
Resumo

CAPÍTULO 3: Metodologia de avaliação

Desafios na avaliação dos modelos de fundação
Entendendo as métricas de modelagem de linguagem
Entropia
Entropia cruzada
Bits-por-caractere e bits-por-byte
Perplexidade
Interpretação e casos de uso da perplexidade
Avaliação exata
Exatidão funcional
Medições de similaridade em relação a dados de referência
Introdução ao embedding
IA como juiz
Por que a IA como juiz?
Como usar a IA como juiz
Limitações da IA como juiz
Quais modelos podem agir como juízes?
Modelos de classificação com avaliação comparativa
Desafios da avaliação comparativa
Futuro da avaliação comparativa
Resumo

CAPÍTULO 4: Avaliação de sistemas de IA

Critérios de avaliação
Capacidade específica do domínio
Capacidade de geração
Capacidade de seguir instruções
Custo e latência
Seleção do modelo
Fluxo de trabalho de seleção de modelos
Construir versus comprar o modelo
Navegue em benchmarks públicos
Projete seu pipeline de avaliação
Etapa 1: Avalie todos os componentes de um sistema
Etapa 2: Crie uma diretriz de avaliação
Etapa 3: Defina métodos e dados de avaliação
Resumo

CAPÍTULO 5: Engenharia de prompt

Introdução ao prompting
Aprendizado no contexto: zero-shot e few-shot
Prompt do sistema e prompt do usuário
Tamanho e eficiência do contexto
Melhores práticas da engenharia de prompt
Escreva instruções claras e explícitas
Forneça contexto suficiente

- Divida tarefas complexas em subtarefas mais simples
- Dê tempo ao modelo para pensar
- Itere em seus prompts
- Avalie ferramentas de engenharia de prompt
- Organize e versione os prompts
- Engenharia de prompt defensiva
- Prompts proprietários e engenharia de prompt reversa
- Jailbreaking e injeção de prompt
- Extração de informações
- Defesas contra ataques de prompt
- Resumo

CAPÍTULO 6: RAG e agentes

- RAG
- Arquitetura RAG
- Algoritmos de recuperação
- Otimização da recuperação
- RAG não é somente para textos
- Agentes
- Visão geral dos agentes
- Ferramentas
- Planejamento
- Modos de falha do agente e avaliação
- Memória
- Resumo

CAPÍTULO 7: Ajuste fino

- Visão geral do ajuste fino
- Quando fazer o ajuste fino
- Razões para realizar o ajuste fino
- Razões para não realizar o ajuste fino
- Ajuste fino e RAG
- Gargalos de memória
- Retropropagação e parâmetros treináveis
- Cálculo de memória
- Representações numéricas
- Quantização
- Técnicas de ajuste fino
- Ajuste fino eficiente dos parâmetros
- Fusão de modelos e ajuste fino multitarefa
- Táticas de ajuste fino
- Resumo

CAPÍTULO 8: Engenharia de dataset

- Curadoria de dados
- Qualidade dos dados
- Cobertura dos dados
- Quantidade dos dados
- Aquisição e anotação de dados

Aumento e síntese dos dados
Por que usar a síntese de dados
Técnicas tradicionais de síntese de dados
Síntese de dados alimentada por IA
Destilação de modelos
Processamento de dados
Inspecione os dados
Desduplique os dados
Limpe e filtre os dados
Formate os dados
Resumo

CAPÍTULO 9: Otimização da inferência
Entendendo a otimização da inferência
Visão geral da inferência
Métricas de desempenho da inferência
Aceleradores de IA
Otimização da inferência
Otimização do modelo
Otimização do serviço de inferência
Resumo

CAPÍTULO 10: Arquitetura da engenharia de IA e feedback do usuário
Arquitetura da engenharia de IA
Etapa 1: melhore o contexto
Etapa 2: insira guardrails
Etapa 3: adicione um roteador de modelo e um gateway
Etapa 4: reduza a latência com caches
Etapa 5: adicione padrões de agentes
Monitoramento e observabilidade
Orquestração de pipeline de IA
Feedback do usuário
Extraindo feedback conversacional
Planejamento do feedback
Limitações do feedback
Resumo

Epílogo

Índice remissivo