Spis treści książki

Wprowadzenie

1. Podstawy modeli GPT-4 i ChatGPT

- Wprowadzenie do modeli LLM
  - Podstawy modeli językowych i NLP
  - Transformer i jego rola w modelu LLM
  - Demistyfikacja etapów tokenizacji i prognozowania w modelach GPT
  - Dodanie wizji do modeli LLM
- Historia modeli w skrócie: od GPT-1 do GPT-4
  - GPT-1
  - GPT-2
  - GPT-3
  - Od GPT-3 do InstructGPT
  - GPT-3.5, Codex i ChatGPT
  - GPT-4
  - Dlaczego używa się systemu ELO do porównywania modeli?
  - Rozwój sztucznej inteligencji w kierunku multimodalności
- Zastosowania modelu LLM i przykładowe produkty
  - Be My Eyes
  - Morgan Stanley
  - Khan Academy
  - Duolingo
  - Yabble
  - Waymark
  - Inworld AI
- Uważaj na halucynacje sztucznej inteligencji: ograniczenia i wnioski
- Uwalnianie potencjału modeli GPT za pomocą zaawansowanych funkcji
- Podsumowanie

2. Szczegółowe informacje o interfejsach OpenAI API

- Podstawowe pojęcia
- Dostępne interfejsy API modeli OpenAI
  - Fundamentalne modele GPT
  - InstructGPT (przestarzałe)
  - GPT-3.5
  - GPT-4
- Testowanie modeli GPT za pomocą platformy OpenAI Playground
- Pierwsze kroki: biblioteka OpenAI dla języka Python
  - Dostęp do modeli i klucz API

        - Przykład "Witaj, świecie!"
    - Korzystanie z modeli uzupełniania rozmów
        - Parametry wejściowe punktu końcowego ChatCompletion
        - Dobieranie wartości parametrów top_p i temperature
        - Format odpowiedzi punktu końcowego ChatCompletion
        - Obraz
        - Wymuszanie odpowiedzi w formacie JSON
    - Korzystanie z innych modeli uzupełniających tekst
        - Parametry wejściowe punktu końcowego Completion
        - Format odpowiedzi punktu końcowego Completion
    - Uwagi
        - Ceny i limity tokenów
        - Bezpieczeństwo i prywatność danych
    - Inne interfejsy API i ich funkcjonalności
        - Osadzenia
        - Tłumaczenie języka naturalnego za pomocą osadzeń w uczeniu maszynowym
        - Modele moderujące
        - Przekształcanie tekstu na mowę
        - Przekształcanie mowy na tekst
        - API do pracy z obrazami
    - Podsumowanie (i ściągawka)

3. Nawigacja po aplikacjach wspieranych przez modele LLM - możliwości i wyzwania

    - Ogólne informacje o tworzeniu aplikacji
        - Zarządzanie kluczami API
        - Bezpieczeństwo i prywatność danych
    - Wzorce architektoniczne oprogramowania
    - Zastosowanie możliwości modeli LLM w aplikacjach
        - Prowadzenie rozmów
        - Przetwarzanie języka
        - Interakcja człowiek - komputer
        - Łączenie zdolności
    - Przykładowe projekty
        - Projekt 1. Generator wiadomości - przetwarzanie języka
        - Projekt 2. Streszczanie filmów z YouTube'a - przetwarzanie języka
        - Projekt 3. Ekspert od Minecrafta - przetwarzanie języka i prowadzenie rozmowy

        - Projekt 4. Osobisty asystent - interfejs interakcji komputer - człowiek
        - Projekt 5. Organizowanie dokumentów - przetwarzanie języka
        - Projekt 6. Analiza wydźwięku tekstu - przetwarzanie języka
    - Zarządzanie kosztami
    - Podatności na ataki aplikacji opartych na modelach LLM
        - Analiza danych wejściowych i wyjściowych
        - Nieuchronność wstrzykiwania promptów
    - Praca z zewnętrznym API
        - Obsługa błędów i nieoczekiwanych opóźnień
        - Limity prędkości
        - Poprawa responsywności i doświadczenia użytkownika
    - Podsumowanie

4. Zaawansowane strategie integracji modeli LLM

- Inżynieria promptu
    - Tworzenie skutecznych promptów z rolami, kontekstem i zadaniami
    - Rozumowanie modelu krok po kroku
    - Implementacja uczenia na kilku przykładach
    - Iteracyjna poprawa z uwzględnieniem opinii użytkownika
    - Zwiększanie skuteczności promptu
- Dostrajanie modelu
    - Pierwsze kroki
    - Dostrajanie modelu za pomocą interfejsu OpenAI API
    - Dostrajanie za pomocą interfejsu webowego OpenAI
    - Zastosowania dostrojonych modeli
    - Generowanie syntetycznych danych i dostrajanie modelu na potrzeby e-mailowej kampanii marketingowej
    - Koszty dostrajania
- RAG
    - Najprostszy RAG
    - Zaawansowany RAG
    - Ograniczenia RAG
- Wybór strategii
    - Porównanie strategii
    - Ocenianie
- Od standardowej aplikacji do rozwiązania wspomaganego przez LLM
    - Wrażliwość na prompt
    - Brak determinizmu
    - Halucynacje

- Podsumowanie

5. Rozszerzanie modeli LLM za pomocą frameworków i wtyczek

- Platforma LangChain
  - Biblioteki platformy Langchain
  - Dynamiczne prompty
  - Agenty i narzędzia
  - Pamięć
  - Osadzenia
- Platforma LlamaIndex
  - Prezentacja: RAG w 10 liniach kodu
  - Zasady platformy LlamaIndex
  - Dostosowanie
- Wtyczki GPT-4
  - Informacje ogólne
  - Interfejs API
  - Manifest
  - Specyfikacja OpenAPI
  - Opisy
- Niestandardowe modele GPT
- API asystentów
  - Tworzenie API asystentów
  - Zarządzanie rozmową z API asystentów
  - Wywoływanie funkcji
  - Asystenty na platformie webowej OpenAI
- Podsumowanie

6. Składanie wszystkiego w całość

- Kluczowe wnioski
- Składanie wszystkiego w całość - przykład zastosowania asystenta
  - Krok 1: tworzenie pomysłów
  - Krok 2: definiowanie wymagań
  - Krok 3: budowanie prototypu
  - Krok 4: ulepszanie, iteracje
  - Krok 5: zabezpieczenie rozwiązania
- Wnioski

Słownik kluczowych pojęć