

Spis treści

Wprowadzenie	9
1. Podstawy modeli GPT-4 i ChatGPT	13
Wprowadzenie do modeli LLM	14
Podstawy modeli językowych i NLP	14
Transformer i jego rola w modelu LLM	17
Demistyfikacja etapów tokenizacji i prognozowania w modelach GPT	20
Dodanie wizji do modeli LLM	22
Historia modeli w skrócie: od GPT-1 do GPT-4	24
GPT-1	24
GPT-2	25
GPT-3	26
Od GPT-3 do InstructGPT	26
GPT-3.5, Codex i ChatGPT	28
GPT-4	30
Rozwój sztucznej inteligencji w kierunku multimodalności	34
Zastosowania modelu LLM i przykładowe produkty	36
Be My Eyes	36
Morgan Stanley	37
Khan Academy	37
Duolingo	38
Yabble	38
Waymark	39
Inworld AI	39
Uważaj na halucynacje sztucznej inteligencji: ograniczenia i wnioski	40
Uwalnianie potencjału modeli GPT za pomocą zaawansowanych funkcji	42
Podsumowanie	46
2. Szczegółowe informacje o interfejsach OpenAI API	48

Podstawowe pojęcia	49
Dostępne interfejsy API modeli OpenAI	50
Fundamentalne modele GPT	50
InstructGPT (przestarzałe)	51
GPT-3.5	51
GPT-4	52
Testowanie modeli GPT za pomocą platformy	
OpenAI Playground	53
Pierwsze kroki: biblioteka OpenAI dla języka Python	57
Dostęp do modeli i klucz API	57
Przykład „Witaj, świecie!”	60
Korzystanie z modeli uzupełniania rozmów	61
Parametry wejściowe punktu końcowego ChatCompletion	63
Dobieranie wartości parametrów top_p i temperature	67
Format odpowiedzi punktu końcowego ChatCompletion	69
Obraz	72
Wymuszanie odpowiedzi w formacie JSON	75
Korzystanie z innych modeli uzupełniających tekst	79
Parametry wejściowe punktu końcowego Completion	80
Format odpowiedzi punktu końcowego Completion	81
Uwagi	82
Ceny i limity tokenów	82
Bezpieczeństwo i prywatność danych	82
Inne interfejsy API i ich funkcjonalności	83
Osadzenia	83
Modele moderujące	86
Przekształcanie tekstu na mowę	90
Przekształcanie mowy na tekst	92
API do pracy z obrazami	95
Podsumowanie (i ściągawka)	106
3. Nawigacja po aplikacjach wspieranych przez modele LLM — możliwości i wyzwania	108

Ogólne informacje o tworzeniu aplikacji	108
Zarządzanie kluczami API	109
Bezpieczeństwo i prywatność danych	111
Wzorce architektoniczne oprogramowania	111
Zastosowanie możliwości modeli LLM w aplikacjach	112
Prowadzenie rozmów	112
Przetwarzanie języka	113
Interakcja człowiek – komputer	115
Łączenie zdolności	116
Przykładowe projekty	117
Projekt 1. Generator wiadomości — przetwarzanie języka	117
Projekt 2. Streszczanie filmów z YouTube’a — przetwarzanie języka	120
Projekt 3. Ekspert od Minecrafta — przetwarzanie języka i prowadzenie rozmowy	124
Projekt 4. Osobisty asystent — interfejs interakcji komputer – człowiek	130
Projekt 5. Organizowanie dokumentów — przetwarzanie języka	138
Projekt 6. Analiza wydźwięku tekstu — przetwarzanie języka	139
Zarządzanie kosztami	146
Podatności na ataki aplikacji opartych na modelach LLM	149
Analiza danych wejściowych i wyjściowych	150
Nieuchronność wstrzykiwania promptów	151
Praca z zewnętrznym API	151
Obsługa błędów i nieoczekiwanych opóźnień	152
Limity prędkości	153
Poprawa responsywności i doświadczenia użytkownika	154
Podsumowanie	158
4. Zaawansowane strategie integracji modeli LLM	159
Inżynieria promptu	159
Tworzenie skutecznych promptów z rolami, kontekstem i zadaniami	160

Rozumowanie modelu krok po kroku	167
Implementacja uczenia na kilku przykładach	169
Iteracyjna poprawa z uwzględnieniem opinii użytkownika	171
Zwiększanie skuteczności promptu	177
Dostrajanie modelu	180
Pierwsze kroki	180
Dostrajanie modelu za pomocą interfejsu OpenAI API	184
Dostrajanie za pomocą interfejsu webowego OpenAI	187
Zastosowania dostrojonych modeli	189
Generowanie syntetycznych danych i dostrajanie modelu na potrzeby e-mailowej kampanii marketingowej	192
Koszty dostrajania	200
RAG	200
Najprostszy RAG	201
Zaawansowany RAG	201
Ograniczenia RAG	208
Wybór strategii	208
Porównanie strategii	209
Ocenianie	212
Od standardowej aplikacji do rozwiązania wspomaganego przez LLM	212
Wrażliwość na prompt	213
Brak determinizmu	213
Halucynacje	214
Podsumowanie	216
5. Rozszerzanie modeli LLM za pomocą frameworków i wtyczek	218
Platforma LangChain	218
Biblioteki platformy Langchain	220
Dynamiczne prompty	221
Agenty i narzędzia	222
Pamięć	226
Osadzenia	228

Platforma LlamaIndex	231
Prezentacja: RAG w 10 liniach kodu	232
Zasady platformy LlamaIndex	232
Dostosowanie	234
Wtyczki GPT-4	236
Informacje ogólne	237
Interfejs API	238
Manifest	239
Specyfikacja OpenAPI	240
Opisy	242
Niestandardowe modele GPT	242
API asystentów	249
Tworzenie API asystentów	250
Zarządzanie rozmową z API asystentów	252
Wywoływanie funkcji	256
Asystenty na platformie webowej OpenAI	260
Podsumowanie	262
6. Składanie wszystkiego w całość	265
Kluczowe wnioski	265
Składanie wszystkiego w całość — przykład zastosowania asystenta	267
Krok 1: tworzenie pomysłów	267
Krok 2: definiowanie wymagań	268
Krok 3: budowanie prototypu	269
Krok 4: ulepszanie, iteracje	269
Krok 5: zabezpieczenie rozwiązania	271
Wnioski	272
Słownik kluczowych pojęć	273
Dodatek. Narzędzia, biblioteki i frameworki	282