

Spis treści

Wstęp	7
-------------	---

Część I. Podstawy	11
--------------------------	-----------

1. Wprowadzenie do inżynierii promptów	13
Modele językowe to magia	14
Modele językowe: Jak do tego doszliśmy?	16
Wczesne modele językowe	18
GPT wkracza na scenę	20
Inżynieria promptów	22
Podsumowanie	24
2. Modele językowe — wprowadzenie	26
Czym są duże modele językowe?	27
Kończenie dokumentu	29
Myślenie człowieka a przetwarzanie danych przez modele językowe	30
Halucynacje	32
Jak modele językowe postrzegają świat?	33
Różnica 1: Modele językowe używają deterministycznej tokenizacji	34
Różnica 2: Modele językowe nie potrafią zwolnić i analizować poszczególnych liter	35
Różnica 3: Modele językowe inaczej postrzegają tekst	37
Zliczanie tokenów	38
Po jednym tokenie na raz	39
Modele autoregresyjne	39
Wzorce i powtórzenia	41
Temperatura i prawdopodobieństwo	42
Architektura transformerów	46
Podsumowanie	52

3. Przejście do czatu	53
Uczenie przez wzmacnianie na podstawie informacji zwrotnych do człowieka	54
Proces budowania modelu RLHF	55
Utrzymywanie rzetelności modeli językowych	58
Unikanie nietypowych zachowań	59
RLHF daje wiele korzyści niewielkim kosztem	59
Uważaj na koszty dostosowania	59
Przejście od instrukcji do konwersacji	60
Modele instrukcyjne	60
Modele konwersacyjne	62
Zmiany w interfejsie API	64
Interfejs API uzupełniania czatu	64
Porównanie konwersacji z uzupełnianiem	67
Od konwersacji do narzędzi	68
Projektowanie promptów jako sztuka dramatopisarska	69
Podsumowanie	71
4. Projektowanie aplikacji LLM	72
Anatomia pętli	72
Problem użytkownika	74
Przekształcanie problemu użytkownika na model dziedziny	75
Użycie LLM do uzupełniania promptu	80
Powrót do dziedziny użytkownika	81
Przyjrzyjmy się bliżej przejściu w przód	82
Tworzenie prostego przejścia w przód	82
Zagadnienie złożoności pętli	84
Ocenianie jakości aplikacji LLM	88
Ocena offline	89
Ocena online	89
Podsumowanie	90

Część II. Podstawowe techniki **91**

5. Treść promptu	93
Źródła treści	94
Treści statyczne	95
Wyjaśnienie zapytania	96
Prompty z kilkoma przykładami	97
Treść dynamiczna	105
Odkrywanie dynamicznego kontekstu	107
Generacja wspomagana wyszukiwaniem	110
Podsumowywanie	121
Podsumowanie	125

6. Konstruowanie promptu	126
Anatomia idealnego promptu	126
Jaki to rodzaj dokumentu?	130
Konwersacja z prośbą o radę	130
Raport analityczny	133
Dokument strukturalny	136
Formatowanie fragmentów	139
Więcej o bezwładności	140
Formatowanie przykładów do promptów	141
Elastyczne fragmenty	142
Powiązania pomiędzy elementami promptów	143
Położenie	143
Ważność	144
Zależność	144
Połączenie wszystkich elementów	146
Podsumowanie	149
7. Okiełznanie modelu	150
Anatomia idealnego uzupełnienia	150
Wstęp	150
Rozpoznawalny początek i koniec	154
Uwaga końcowa	155
Nie tylko tekst: Logarytmy prawdopodobieństw	156
Jak dobra jest generowana treść?	157
Stosowanie modeli LLM do klasyfikacji	158
Kluczowe miejsca promptu	160
Wybór modelu	162
Podsumowanie	169

Część III. Ekspert sztuki **171**

8. Sprawczość konwersacyjna	173
Stosowanie narzędzi	173
Modele LLM przystosowane do korzystania z narzędzi	174
Wytyczne dotyczące definiowania narzędzi	182
Rozumowanie	184
Rozumowanie krok po kroku	185
ReAct: Interaktywne rozumowanie i działanie	186
Nie tylko ReAct	189
Kontekst interakcji bazujących na zadaniach	190
Źródła kontekstu	190
Wybór i organizacja kontekstu	192

Tworzenie agenta konwersacyjnego	194
Zarządzanie konwersacjami	194
Doświadczenia użytkownika	197
Podsumowanie	199
9. Przepływy pracy korzystające z modeli LLM	201
Czy interfejs konwersacyjny wystarczy?	202
Podstawowe przepływy pracy korzystające z LLM	206
Zadania	207
Tworzenie przepływu pracy	213
Przykładowy przepływ pracy: Marketing wtyczek do Shopify	216
Zaawansowane przepływy pracy korzystające z modeli LLM	219
Umożliwienie agentowi LLM sterowania przepływem pracy	220
Agenty zadaniowe z pamięcią stanu	221
Role i delegacje	222
Podsumowanie	223
10. Ocena aplikacji korzystających z modeli LLM	225
Co właściwie testujemy?	226
Ocenianie offline	227
Przykłady zestawów testowych	227
Poszukiwanie próbek	230
Ocenianie rozwiązań	232
Oceny SOMA	236
Testy online	240
Testy A/B	240
Metryki	241
Podsumowanie	244
11. Rzut oka w przyszłość	245
Multimodalność	245
Doświadczenie użytkownika i interfejs użytkownika	246
Inteligencja	248
Podsumowanie	250