

Spis treści

Wprowadzenie 11

1. Pięć zasad promptowania 17

1. Określ wytyczne 23

2. Określ format odpowiedzi 28

3. Przedstaw przykłady 30

4. Oceniaj jakość 34

5. Dziel pracę 45

Podsumowanie 50

2. Wprowadzenie do dużych modeli językowych do generowania tekstu 52

Czym są modele do generowania tekstu? 52

Reprezentacje wektorowe: język w liczbach 53

Architektura transformerów: orkiestracja związków kontekstowych 54

Probabilistyczne generowanie tekstu: mechanizm podejmowania decyzji 55

Krótko o historii: wzrost znaczenia architektur transformerów 56

Wstępnie przeszkolony transformer generatywny OpenAI 58

GPT-3.5-turbo i ChatGPT 59

GPT-4 61

Google Gemini 61

Model Llama firmy Meta a otwarte oprogramowanie 62

Kwantyzacja i LoRA 63

Mistral 63

Anthropic: Claude 64

GPT-4V(ision) 64

Porównanie modeli 64

Podsumowanie 663. Standardowe zasady generowania tekstu z ChatGPT 67

Generowanie list 67

Generowanie list hierarchicznych 69

Kiedy unikać wyrażeń regularnych? 73

Generowanie danych w formacie JSON 73

Generowanie danych w formacie YAML 75

Filtrowanie dokumentów YAML 76

Obsługa nieprawidłowych dokumentów w formacie YAML 77

Generowanie różnorodnych formatów z użyciem ChatGPT 80

Spreparowane dane CSV 80

Wyjaśnij to jak pięciolatkowi 81

Uniwersalne tłumaczenia za pomocą LLM 82

Pytaj o kontekst 84

Wydzielenie stylu tekstu 87

Znalezienie pożądanych cech tekstu 87

Generowanie nowej treści za pomocą wyekstrahowanych cech 89

Ekstrakcja określonych cech tekstu za pomocą LLM 89

Podsumowywanie 89

Podsumowywanie a ograniczenia okna kontekstu 90

Podział tekstu na fragmenty 92

Zalety dzielenia tekstu na fragmenty 92

Scenariusze podziału tekstu na fragmenty 93

Niewłaściwe przykłady dzielenia tekstu na fragmenty 93

Strategie podziału 95

Wykrywanie zdań z użyciem spaCy 96

Tworzenie prostego algorytmu podziału w Pythonie 97

Podział za pomocą okna przesuwającego 98

Pakiety do dzielenia tekstu 100

Podział tekstu z biblioteką tiktoken 100

Kodowania 101

Zrozumienie procesu tokenizacji łańcuchów znaków 101

Szacowanie użycia tokenów w wywołaniach API czata 102

Analiza sentymentu 104

Sposoby usprawnienia analizy sentymentu 105

Ograniczenia i wyzwania analizy sentymentu 106

Od najmniejszych do największych 106

Planowanie architektury 107

Programowanie funkcji we Flasku 107

Dodawanie testów	108
Zalety techniki od najmniejszych do największych	108
Wyzwania, jakie wiążą się z techniką od najmniejszych do największych	109
Promptowanie z użyciem ról	109
Zalety promptowania z użyciem ról	110
Wyzwania, jakie wiążą się z promptowaniem z użyciem ról	111
Kiedy korzystać z promptowania z użyciem ról	111
Techniki promptowania GPT	112
Unikanie halucynacji dzięki tekstom źródłowym	112
Daj modelowi GPT „czas na przemyślenie”	113
Technika wewnętrznego monologu	114
Samooceniające odpowiedzi modelu językowego	115
Klasyfikacja za pomocą dużych modeli językowych	116
Tworzenie modelu klasyfikacji	117
Głosowanie większością w klasyfikacji	118
Ewaluacja kryteriów	119
Metapromptowanie	122
Podsumowanie	125
4. Zaawansowane techniki generowania tekstu za pomocą LangChain	126
Wprowadzenie do LangChain	126
Konfiguracja środowiska	128
Modele czatowe	129
Strumieniowanie modeli czatowych	130
Generowanie wielu odpowiedzi z dużych modeli językowych	131
Szablony promptów LangChain	131
Język wyrażeń LangChain (LCEL)	132
Stosowanie PromptTemplate z modelami czatowymi	134
Parsery wyjścia	134
Ewaluacje LangChain	138
Wywołanie funkcji OpenAI	145
Współbieżne wywołanie funkcji	148
Wywoływanie funkcji w LangChain	149

Ekstrakcja danych za pomocą LangChain	151
Planowanie zapytań	151
Tworzenie szablonów promptów z kilkoma przykładami	153
Podejście z kilkoma przykładami o stałej długości	153
Formatowanie przykładów	154
Wybór promptów z kilkoma przykładami według długości	155
Ograniczenia promptów z kilkoma przykładami	157
Zapisywanie i wczytywanie promptów modeli językowych	157
Łączenie danych	158
Ładowarki dokumentów	160
Rozdzielacze tekstu	162
Poleć książkęKup książkę	
8 Spis treści	
Podział tekstu według długości i liczby tokenów	163
Rekursywny podział według wielu znaków	164
Dekompozycja zadań	166
Łącuchy promptów	168
Łącuchy sekwencyjne	168
Ekstrakcja kluczy za pomocą funkcji itemgetter	170
Tworzenie struktury dla łańcuchów LCEL	175
Łącuchy dokumentów	175
Łącuch nadziania dokumentów	177
Łącuchy oczyszczania	178
Mapowanie i redukcja	178
Łącuch mapowania z rankingiem	179
Podsumowanie	180
5. Wektorowe bazy danych z FAISS i Pinecone	181
Retrieval Augmented Generation (RAG)	184
Wprowadzenie do osadzeń	185
Ładowanie dokumentów	193
Pozyskiwanie z pamięci za pomocą FAISS	195
RAG z użyciem frameworka LangChain	199
Wektorowe bazy danych w chmurze z użyciem Pinecone	201

Samoodpytywanie 208

Alternatywne mechanizmy pozyskiwania danych 212

Podsumowanie 214

6. Agenty autonomiczne z pamięcią i narzędziami 215

Łańcuch myśli 215

Agenty 217

Wnioskuj i działaj (ReAct) 219

Implementacja schematu wnioskuj i działaj 221

Stosowanie narzędzi 227

Duże modele językowe jako API (funkcje OpenAI) 228

Porównanie funkcji OpenAI i schematu ReAct 232

Przypadki użycia dla funkcji OpenAI 233

ReAct 233

Przypadki użycia dla schematu ReAct 234

Zestawy narzędzi dla agentów 234

Dostosowywanie standardowych agentów 236

Własne agenty w LCEL 237

Zasady użycia pamięci 239

Pamięć długoterminowa 239

Pamięć krótkoterminowa 240

Pamięć krótkoterminowa w agentach konwersacyjnych QA 240

Obsługa pamięci w LangChain 241

Zachowywanie stanu 242

Odpytywanie stanu 242

ConversationBufferMemory 242

Inne popularne rodzaje pamięci w LangChain 245

ConversationBufferWindowMemory 245

ConversationSummaryMemory 245

ConversationSummaryBufferMemory 246

ConversationTokenBufferMemory 246

Agent funkcji OpenAI z pamięcią 247

Zaawansowane frameworki agentowe 249

Agenty typu planuj i uruchamiaj 249

Drzewo myśli 250

Wywołania zwrotne 251

Globalne wywołania zwrotne (w konstruktorach) 253

Wywołania zwrotne żądań 253

Argument verbose 254

Które wywołanie zwrotne wybrać? 254

Zliczanie tokenów za pomocą LangChain 254

Podsumowanie 256

7. Wprowadzenie do modeli dyfuzyjnych przeznaczonych
do generowania obrazów 257

OpenAI DALL-E 260

Midjourney 262

Stable Diffusion 265

Google Gemini 268

Generowanie filmów na podstawie tekstu 268

Porównanie modeli 268

Podsumowanie 269

8. Standardowe metody generowania obrazów z Midjourney 270

Modyfikatory formatu 270

Modyfikatory stylu w sztuce 274

Inżynieria odwrotna promptów 276

Dopalacze jakości 276

Prompty negatywne 278

Pojęcia ważne 281

Promptowanie z użyciem obrazków 284

Wmalowywanie 286

Domalowywanie 288

Spójne postaci 290

Przepisywanie promptów 292

Rozdzielenie memów 294

Mapowanie memów 298

Analiza promptów 300

Podsumowanie 302

9. Zaawansowane techniki generowania obrazów za pomocą Stable Diffusion 303

Uruchamianie modelu Stable Diffusion 303

Interfejs webowy AUTOMATIC1111 309

Img2Img 316

Skalowanie obrazków w górę 319

Tryb Interrogate CLIP 322

Wmalowanie i domalowanie w Stable Diffusion 322

ControlNet 325

Model segmentowania wszystkiego (SAM) 334

Dostrajanie DreamBooth 336

Dostrajacz modelu Stable Diffusion XL 343

Podsumowanie 346

10. Tworzenie aplikacji wspomaganych AI 348

Pisanie bloga przez AI 348

Badanie tematu 349

Wywiad ekspercki 352

Wygeneruj zarys 353

Generowanie tekstu 355

Styl pisanie 357

Optymalizacja tytułu 360

Obrazki na blogu generowane przez AI 361

Interfejs użytkownika 366

Podsumowanie 368

Skorowidz 370