

Spis treści

O autorze	9
O korektorach merytorycznych	10
Wprowadzenie	11
ROZDZIAŁ 1.	
Wprowadzenie do dużych modeli językowych	19
Czym jest duży model podstawowy i duży model językowy?	20
Przesunięcie paradygmatu sztucznej inteligencji — wprowadzenie do modeli podstawowych	21
Budowa dużego modelu językowego	23
Najpopularniejsze architektury oparte na transformerach dużego modelu językowego	30
Wczesne eksperymenty	30
Wprowadzenie do architektury transformera	31
Trenowanie dużego modelu językowego i jego ocena	37
Trenowanie dużego modelu językowego	37
Ocena modelu	40
Modele podstawowe kontra modele dostosowane do własnych potrzeb	43
Jak można dostosować model do własnych potrzeb?	43
Podsumowanie	46
Odwołania	46
ROZDZIAŁ 2.	
Duże modele językowe w aplikacjach wspomaganych przez sztuczną inteligencję	47
Sposoby, w jakie duże modele językowe zmieniły branżę tworzenia oprogramowania	48
System copilot	49
Wprowadzenie do frameworków koordynowania sztucznej inteligencji umożliwiających osadzanie dużych modeli językowych w aplikacjach	54

Najważniejsze komponenty frameworków koordynowania	
sztucznej inteligencji	54
LangChain	57
Haystack	59
Semantic Kernel	60
Jak wybrać framework?	63
Podsumowanie	65
Odwołania	65
ROZDZIAŁ 3.	
Wybór dużego modelu językowego dla aplikacji	66
Ogólne omówienie najbardziej obiecujących	
dużych modeli językowych dostępnych na rynku	66
Modele własnościowe	67
Modele otwartoźródłowe	77
Nie tylko modele językowe	82
Kryteria stosowane podczas wyboru właściwego	
dużego modelu językowego	85
Rozważania	88
Studium przypadku	90
Podsumowanie	91
Odwołania	92
ROZDZIAŁ 4.	
Prompt engineering	93
Wymagania techniczne	93
Czym jest prompt engineering?	94
Reguły prompt engineering	94
Jasne instrukcje	95
Podziel skomplikowane zadania na mniejsze	97
Zapytaj o uzasadnienie	100
Wygenerowanie wielu danych wyjściowych,	
a następnie użycie modelu do wybrania tych najlepszych	102

Powtórzenie instrukcji na końcu	104
Używanie ograniczników	106
Techniki zaawansowane	109
Technika few-shots	109
Technika chain of thought	112
ReAct	114
Podsumowanie	118
Odwołania	118
ROZDZIAŁ 5.	
Osadzanie dużych modeli językowych w aplikacjach	120
Wymagania techniczne	120
Wprowadzenie do frameworka LangChain	121
Rozpoczęcie pracy z frameworkiem LangChain	123
Modele i prompty	123
Połączenia danych	126
Pamięć	132
Łańcuch	135
Agenty	140
Praca z dużymi modelami językowymi poprzez platformę Hugging Face Hub	143
Tworzenie tokena dostępu użytkownika Hugging Face Hub	143
Przechowywanie kluczy tajnych użytkownika w pliku .env	146
Rozpoczęcie pracy z otwartoźródłowymi dużymi modelami językowymi	147
Podsumowanie	148
Odwołania	149
ROZDZIAŁ 6.	
Tworzenie aplikacji konwersacyjnych	150
Wymagania techniczne	150
Rozpoczęcie pracy z aplikacjami konwersacyjnymi	151
Tworzenie zwykłego bota	152

Dodawanie pamięci	153
Dodawanie wiedzy nieparametrycznej	157
Dodawanie narzędzi zewnętrznych	160
Opracowanie frontendu za pomocą Streamlit	163
Podsumowanie	167
Odwołania	168

ROZDZIAŁ 7.

Używanie dużych modeli językowych do tworzenia silników

wyszukiwania i rekomendacji	169
Wymagania techniczne	169
Wprowadzenie do systemu rekomendacji	170
Istniejące systemy rekomendacji	171
Algorytm K najbliższych sąsiadów	172
Rozkład macierzy	173
Sieci neuronowe	176
Jak duże modele językowe zmieniają systemy rekomendacji?	179
Implementowanie systemu rekomendacji	
opartego na dużym modelu językowym	180
Wstępne przygotowanie danych	181
Tworzenie chatbota rekomendacji w scenariuszu typu zimny start	184
Tworzenie systemu opartego na treści	191
Opracowanie frontendu za pomocą biblioteki Streamlit	196
Podsumowanie	198
Odwołania	199

ROZDZIAŁ 8.

Używanie dużych modeli językowych w połączeniu

z danymi strukturyzowanymi	200
Wymagania techniczne	200
Czym są dane strukturyzowane?	201
Rozpoczęcie pracy z relacyjnymi bazami danych	203
Wprowadzenie do relacyjnych baz danych	203

Ogólne omówienie bazy danych Chinook	205
Jak za pomocą Pythona pracować z relacyjną bazą danych?	207
Implementacja aplikacji DBCopilot z użyciem frameworka LangChain	211
Agenty frameworka LangChain i agent SQL	211
Prompt engineering	215
Dodawanie kolejnych narzędzi	218
Opracowanie frontendu za pomocą biblioteki Streamlit	222
Podsumowanie	225
Odwołania	226
ROZDZIAŁ 9.	
Praca z kodem źródłowym	227
Wymagania techniczne	227
Wybór odpowiedniego dużego modelu językowego przeznaczonego do pracy z kodem źródłowym	228
Analizowanie i generowanie kodu źródłowego	230
Falcon LLM	231
CodeLlama	234
StarCoder	238
Działanie w charakterze algorytmu	243
Wykorzystanie API Code Interpreter	250
Podsumowanie	256
Odwołania	256
ROZDZIAŁ 10.	
Tworzenie wielomodalnych aplikacji wykorzystujących duże modele językowe	258
Wymagania techniczne	258
Dlaczego wielomodalność?	259
Tworzenie agenta wielomodalnego za pomocą frameworka LangChain	260
Opcja 1. Użycie standardowego zbioru narzędzi dla Azure AI Services	262

Rozpoczęcie pracy z AzureCognitiveServicesToolkit	262
Opcja 2. Połączenie narzędzi w postać pojedynczego agenta	276
Narzędzia YouTube i Whisper	277
DALL-E i generowanie tekstu	279
Połączenie wszystkiego w całość	281
Opcja 3. Podejście na stałe zdefiniowane w kodzie	
z użyciem łańcucha sekwencyjnego	285
Porównanie wszystkich trzech opcji	288
Opracowanie frontendu za pomocą biblioteki Streamlit	290
Podsumowanie	292
Odwołania	293
ROZDZIAŁ 11.	
Dostrajanie dużych modeli językowych	294
Wymagania techniczne	295
Czym jest dostrajanie?	295
Kiedy dostrajanie modelu jest konieczne?	298
Rozpoczęcie dostrajania	300
Pobranie zbioru danych	300
Tokenizacja danych	302
Dostrajanie modelu	304
Używanie wskaźników oceny modelu	306
Trenowanie i zapisywanie	310
Podsumowanie	313
Odwołania	313
ROZDZIAŁ 12.	
Odpowiedzialna sztuczna inteligencja	314
Czym jest odpowiedzialna sztuczna inteligencja	
i dlaczego jej potrzebujemy?	314
Architektura odpowiedzialnej sztucznej inteligencji	316
Warstwa modelu	317
Warstwa metapromptu	320

Warstwa interfejsu użytkownika	321
Regulacje prawne dotyczące odpowiedzialnej sztucznej inteligencji	325
Podsumowanie	327
Odwołania	327
ROZDZIAŁ 13.	
Najnowsze trendy i innowacje	329
Najnowsze trendy w modelach językowych i generatywnej sztucznej inteligencji	329
GPT-4V(ision)	330
DALL-E 3	332
AutoGen	332
Małe modele językowe	334
Firmy wykorzystujące generatywną sztuczną inteligencję	335
Coca-Cola	335
Notion	336
Malbek	337
Microsoft	337
Podsumowanie	338
Odwołania	340