
Spis treści

Wstęp	11
1. Wprowadzenie do tworzenia aplikacji AI przy użyciu modeli podstawowych	21
Rozwój inżynierii AI	22
Od modeli językowych do dużych modeli językowych	22
Od dużych modeli językowych do modeli podstawowych	28
Od modeli podstawowych do inżynierii AI	31
Przykłady zastosowań modeli podstawowych	34
Programowanie	39
Tworzenie obrazów i wideo	41
Generowanie tekstów	41
Edukacja	43
Boty konwersacyjne	44
Agregacja informacji	45
Organizowanie danych	46
Automatyzacja przepływu danych	46
Planowanie aplikacji AI	47
Ocena przypadków użycia	47
Ustalenie oczekiwań	51
Zaplanowanie kamieni milowych	52
Konserwacja i utrzymanie	52
Stos technologiczny inżynierii AI	54
Trzy warstwy stosu AI	55
Inżynieria AI a inżynieria ML	57
Inżynieria AI a inżynieria pełnowymiarowa	64
Podsumowanie	65

2. Zrozumienie modeli podstawowych	67
Dane treningowe	68
Modele wielojęzyczne	69
Modele specyficzne dla danej dziedziny	73
Modelowanie	75
Architektura modelu	75
Rozmiar modelu	84
Post-trening	93
Dostrajanie nadzorowane	96
Dostrajanie preferencji	99
Próbkowanie	103
Podstawy próbkowania	103
Strategie próbkowania	105
Efektywność obliczeniowa czasu testowania	110
Dane ustrukturyzowane	113
Probabilistyczna natura sztucznej inteligencji	119
Podsumowanie	125
3. Metodyka ewaluacji	127
Wyzwania związane z ewaluacją modeli podstawowych	128
Zrozumienie wskaźników dotyczących modelowania języka	132
Entropia	133
Entropia krzyżowa	134
Wskaźniki bits-per-character i bits-per-byte	134
Nieokreśloność	135
Interpretacja nieokreśloności i jej zastosowania	136
Ewaluacja dokładna	138
Poprawność funkcjonalna	139
Pomiar poziomu podobieństwa względem danych referencyjnych	140
Wprowadzenie do osadzania	146
AI jako sędzia	148
Dlaczego „AI jako sędzia”?	149
Jak używać metody „AI jako sędzia”?	150
Ograniczenia metody „AI jako sędzia”	153
Jakie modele mogą być sędziami?	156
Ranking modeli wynikający z ewaluacji porównawczej	159
Wyzwania ewaluacji porównawczej	163
Przyszłość ewaluacji porównawczej	166
Podsumowanie	167

4. Ewaluacja modeli AI	168
Kryteria ewaluacji	168
Zdolności specyficzne dla danej dziedziny	170
Zdolności generacyjne	172
Zdolność do podążania za instrukcjami	180
Koszty i opóźnienia	185
Wybór modelu	187
Proces wyboru modelu	187
Budowa czy zakup modelu?	189
Zestawy testów dostępne publicznie	199
Projektowanie procesu ewaluacji	208
Krok 1. Ewaluacja wszystkich komponentów systemu	208
Krok 2. Utworzenie wytycznych do ewaluacji	210
Krok 3. Określenie metod ewaluacji i danych	212
Podsumowanie	216
5. Inżynieria promptów	218
Wprowadzenie do tworzenia promptów	219
Uczenie w kontekście: zero-shot i few-shot	220
Prompt systemowy a prompt użytkownika	222
Długość i efektywność kontekstu	224
Najlepsze zasady inżynierii promptów	226
Twórz jasno i precyzyjnie sformułowane instrukcje	227
Dostarcz niezbędny kontekst	229
Podziel zadania złożone na prostsze podzadania	231
Daj modelowi czas na myślenie	233
Doskonal prompty w procesie iteracyjnym	234
Oceniaj narzędzia do inżynierii promptów	235
Porządkuj prompty i zarządzaj ich wersjami	238
Strategia zabezpieczania promptów	240
Prompty zastrzeżone i inżynieria odwrotna promptów	242
Omijanie zabezpieczeń i wstrzykiwanie promptów	243
Ekstrakcja informacji	248
Obrona przed atakami na prompty	252
Podsumowanie	256
6. Generowanie wspomagane wyszukiwaniem i agenty	257
Generowanie wspomagane wyszukiwaniem	258
Architektura generowania wspomaganego wyszukiwaniem	260
Algorytmy wyszukiwania	261
Optymalizacja wyszukiwania	272
Generowanie wspomagane wyszukiwaniem a inne modalności	277

Agenty	280
Przegląd agentów	281
Narzędzia	283
Planowanie	286
Tryby błędów agenta i sposoby ich oceny	301
Pamięć	304
Podsumowanie	308
7. Dostrajanie	310
Wprowadzenie do dostrajania	311
Kiedy należy dostrajać?	314
Powody, by dostrajać	314
Powody, by nie dostrajać	315
Dostrajanie a generowanie wspomagane wyszukiwaniem (RAG)	319
Ograniczenia pamięciowe	322
Propagacja wsteczna i parametry trenowane	323
Obliczenia dotyczące pamięci	325
Reprezentacje numeryczne	328
Kwantyzacja	330
Techniki dostrajania	334
Dostrajanie efektywne parametrowo	335
Scalanie modeli i dostrajanie wielozadaniowe	348
Taktyki dostrajania	357
Podsumowanie	361
8. Inżynieria zbiorów danych	363
Przygotowanie danych	365
Jakość danych	367
Pokrycie danych	369
Ilość danych	371
Pozyskiwanie i etykietowanie danych	376
Synteza i generowanie sztucznych danych	379
Po co stosować syntezę danych?	379
Tradycyjne metody syntezy danych	381
Synteza danych wspierana przez AI	385
Destylacja modelu	393
Przetwarzanie danych	394
Inspekcja danych	395
Deduplikacja danych	396
Czyszczenie i filtrowanie danych	398
Formatowanie danych	399
Podsumowanie	400

9. Optymalizacja wnioskowania	402
Zrozumienie optymalizacji wnioskowania	403
Podstawy wnioskowania	403
Wskaźniki związane z wydajnością wnioskowania	409
Akceleratory AI	415
Optymalizacja wnioskowania	422
Optymalizacja modelu	422
Optymalizacja usługi wnioskowania	436
Podsumowanie	443
10. Architektura systemów AI i informacje zwrotne od użytkowników	445
Architektura systemów AI	445
Krok 1. Rozszerzenie kontekstu	446
Krok 2. Wprowadzenie zabezpieczeń	447
Krok 3. Wprowadzenie routingu i bramki dostępowej	451
Krok 4. Zmniejszenie opóźnień za pomocą mechanizmów buforowania	456
Krok 5. Dodanie wzorców agentowych	459
Monitorowanie systemu i przejrzystość jego działania	461
Orkiestracja potoku AI	467
Informacja zwrotna od użytkowników	469
Pozyskiwanie informacji zwrotnej z rozmów	469
Projektowanie systemu gromadzenia informacji zwrotnych	475
Ograniczenia systemu gromadzenia informacji zwrotnych	482
Podsumowanie	485
Epilog	487
Skorowidz	488