

Francesco Esposito

Programowanie wielkich modeli językowych z użyciem Azure Open AI: Programowanie konwersacyjne i inżynieria odpowiedzi z wykorzystaniem modeli LLM

Przekład: Krzysztof Kapustka

APN Promise, Warszawa 2024

Spis treści

<i>Podziękowania</i>	ix
<i>Wprowadzenie</i>	xi
1 Geneza i analiza wielkich modeli językowych	1
Pierwsze spojrzenie na modele LLM	1
Historia modeli LLM	2
Podstawy funkcjonowania	7
Biznesowe przypadki użycia	17
Fakty dotyczące programowania konwersacyjnego	18
Rodząca się moc języka naturalnego	18
Topologia modeli LLM	20
Perspektywa przyszłości	23
Podsumowanie	29
2 Podstawowe techniki uczenia z użyciem podpowiedzi	31
Czym jest inżynieria podpowiedzi?	31
Pierwsze spojrzenie na podpowiedzi	32
Alternatywne sposoby na zmianę wyniku	35
Przygotowanie do wykonywania kodu	38
Techniki podstawowe	43
Scenariusze zero-shot	43
Scenariusze few-shot	45
Scenariusze łańcucha myśli	50
Podstawowe przypadki użycia	54
Chatboty	55
Tłumaczenie	58
Ograniczenia modeli LLM	59
Podsumowanie	60
3 Projektowanie zaawansowanych podpowiedzi	61
Co poza inżynierią podpowiedzi?	62
Łączenie elementów	62
Dostrajanie	65
Wywoływanie funkcji	68
Styl ręczny	68

Styl OpenAI	72
Rozmawianie z (odseparowanymi) danymi	77
Łączenie danych z modelem LLM	78
Osadzenia	78
Magazyn wektorów	84
Generowanie wzbogacone wyszukiwaniem informacji	87
Podsumowanie	92
4 Korzystanie z platform językowych	95
Potrzeba orkiestratora	95
Koncepcje międzyplatformowe	97
Punkty do rozważenia	103
LangChain	106
Modele, szablony podpowiedzi i łańcuchy	106
Agenci	115
Połączenie danych	125
Microsoft Semantic Kernel	131
Wtyczki	132
Dane i planiści	138
Microsoft Guidance	144
Konfiguracja	144
Główne funkcje	147
Podsumowanie	153
5 Obawy związane z bezpieczeństwem, prywatnością i dokładnością	155
Omówienie	155
Odpowiedzialna AI	156
Red teaming	157
Filtrowanie nadużyć i treści	158
Halucynacja i efektywność	159
Stronniczość i bezstronność	161
Bezpieczeństwo i prywatność	161
Bezpieczeństwo	161
Prywatność	167
Ocena i filtrowanie treści	173
Ewaluacja	173
Filtrowanie treści	178
Podsumowanie	188
6 Budowanie osobistego asystenta	189
Omówienie aplikacji internetowej chatbota	190

Zakres	190
Stos technologiczny	191
Projekt	192
Konfiguracja modelu LLM	192
Konfigurowanie projektu	194
Integrowanie modelu LLM	196
Możliwe rozszerzenia	210
Podsumowanie	212
7 Czatowanie z naszymi danymi	213
Omówienie	213
Zakres	213
Stos technologiczny	214
Co to jest Streamlit?	215
Krótkie wprowadzenie do Streamlit	215
Główne funkcje interfejsu użytkownika	216
Plusy i minusy w produkcji	218
Projekt	219
Konfigurowanie projektu i podstawowego interfejsu użytkownika	220
Przygotowywanie danych	223
Integracja z modelem LLM	228
Dalszy postęp	233
Generowanie wspomagane wyszukiwaniem informacji kontra dostrajanie	233
Możliwe rozszerzenia	236
Podsumowanie	237
8 Konwersacyjny interfejs użytkownika	239
Omówienie	240
Zakres	240
Stos technologiczny	241
Projekt	243
Konfiguracja API Minimal	243
OpenAPI	245
Integracja z modelem LLM	247
Możliwe rozszerzenia	254
Podsumowanie	255
A Wewnętrzne funkcjonowanie modeli LLM	257
Rola prawdopodobieństwa	257
Podejście heurystyczne	258
Sztuczne neurony	260

Przypadek modelu GPT	268
Transformator i uwaga.....	269
Szkolenia i nowe możliwości.....	274
<i>Indeks</i>	279