

Spis treści

O autorze	17
O korektorze merytorycznym	18
Przedmowa	19
ROZDZIAŁ 1.	
Czym są transformery?	27
Stała złożoność czasowa $O(1)$, która na zawsze zmieniła nasze życie	29
Uwaga $O(1)$ pokonuje rekurencyjne metody $O(n)$	30
Magia obliczeniowej złożoności czasowej warstwy uwagi	32
Krótką podróż od rekurencji do uwagi	41
Od jednego tokena do rewolucji w dziedzinie sztucznej inteligencji	44
Od jednego tokena do całości	47
Modele podstawowe	49
Od zadań ogólnych do zadań specjalistycznych	50
Rola specjalistów AI	55
Przyszłość specjalistów AI	56
Jakich zasobów powinniśmy używać?	57
Wytyczne dotyczące podejmowania decyzji	58
Rozwój łatwych do integracji interfejsów API i asystentów	59
Wybieranie gotowych do użycia bibliotek opartych na API	61
Wybór platformy chmurowej i modelu transformera	62
Podsumowanie	63
Pytania	64
Odnośniki	64
Lektura uzupełniająca	65
ROZDZIAŁ 2.	
Wprowadzenie do architektury modelu transformera	66
Powstanie transformera — uwaga to wszystko, czego potrzebujesz	67
Stos kodera	70
Stos dekodera	93

Szkolenie i wydajność	96
Transformery Hugging Face	97
Podsumowanie	98
Pytania	99
Odnośniki	100
Lektura uzupełniająca	100

ROZDZIAŁ 3.

Emergencja a zadania końcowe — niewidoczne głębiny transformerów	101
Zmiana paradygmatu: czym jest zadanie NLP?	102
Wewnątrz głowicy podwarstwy uwagi transformera	103
Analiza emergencji z użyciem ChatGPT	107
Badanie potencjału modelu w zakresie wykonywania zadań końcowych	110
Ocena modeli za pomocą wskaźników	111
Ocena dokonywana przez człowieka	112
Uruchamianie zadań końcowych	124
CoLA	124
SST-2	125
MRPC	126
WSC	127
Podsumowanie	127
Pytania	128
Odnośniki	129
Lektura uzupełniająca	130

ROZDZIAŁ 4.

Postępy w tłumaczeniach z wykorzystaniem Google Trax, Tłumacza Google i Gemini	131
Definicja tłumaczenia maszynowego	132
Transdukcje i tłumaczenia wykonywane przez ludzi	133
Transdukcje i tłumaczenia maszynowe	134

Ocena tłumaczeń maszynowych	135
Wstępne przetwarzanie zbioru danych WMT	135
Ocena tłumaczeń maszynowych według BLEU	141
Tłumaczenia z wykorzystaniem Google Trax	144
Instalowanie biblioteki Trax	145
Tworzenie modelu oryginalnego transformera	145
Inicjalizowanie modelu z wykorzystaniem wyuczonych wag	147
Tokenizowanie zdania	147
Dekodowanie wyjścia z transformera	147
Detokenizowanie i wyświetlanie tłumaczenia	148
Tłumaczenie za pomocą Tłumacza Google	149
Tłumaczenie z wykorzystaniem wrappera interfejsu	
Google Translate Ajax API	150
Tłumaczenie z wykorzystaniem systemu Gemini	152
Potencjał systemu Gemini	153
Podsumowanie	154
Pytania	155
Odnośniki	155
Lektura uzupełniająca	156
ROZDZIAŁ 5.	
Szczegóły dostrajania z wykorzystaniem modelu BERT	157
Architektura BERT	158
Stos kodera	159
Dostrajanie modelu BERT	166
Określanie celu	167
Ograniczenia sprzętowe	167
Instalowanie transformerów Hugging Face	168
Importowanie modułów	168
Określanie CUDA jako urządzenia dla modułu torch	169
Ładowanie zestawu danych CoLA	169
Tworzenie zdań i list etykiet oraz dodawanie tokenów BERT	171

Aktywowanie tokenizera BERT	171
Przetwarzanie danych	172
Tworzenie masek uwagi	172
Dzielenie danych na zbiór szkoleniowy i zbiór walidacyjny	173
Konwertowanie danych na tensory torch	173
Wybieranie rozmiaru partii i tworzenie iteratora	173
Konfigurowanie modelu BERT	174
Ładowanie bazowego modelu Hugging Face bert-base-uncased	176
Pogrupowane parametry optymalizatora	177
Hiperparametry pętli szkoleniowej	179
Pętla szkolenia	180
Ocena szkolenia	181
Prognozowanie i ocena z użyciem wydzielonego zbioru danych	181
Ocena modelu z wykorzystaniem współczynnika korelacji Matthews'a	184
Ocena za pomocą współczynnika korelacji Matthews'a całego zestawu danych	185
Budowanie interfejsu Pythona do interakcji z modelem	186
Zapisywanie modelu	186
Tworzenie interfejsu dla przeszkolonego modelu	187
Podsumowanie	189
Pytania	190
Odnosińki	191
Lektura uzupełniająca	191
ROZDZIAŁ 6.	
Wstępne szkolenie transformera od podstaw z wykorzystaniem modelu RoBERTa	192
Szkolenie tokenizera i wstępne szkolenie transformera	194
Budowanie modelu KantaiBERT od podstaw	195
Krok 1. Ładowanie zbioru danych	196
Krok 2. Instalowanie transformerów Hugging Face	197

Krok 3. Szkolenie tokenizera	198
Krok 4. Zapisywanie plików na dysku	200
Krok 5. Ładowanie plików tokenizera po przeszkoleniu	201
Krok 6. Sprawdzanie ograniczeń zasobów: GPU i CUDA	202
Krok 7. Definiowanie konfiguracji modelu	203
Krok 8. Ponowne ładowanie tokenizera w module transformers	203
Krok 9. Inicjalizowanie modelu od podstaw	203
Krok 10. Tworzenie zbioru danych	209
Krok 11. Definiowanie mechanizmu zbierania danych	210
Krok 12. Inicjalizowanie trenera	210
Krok 13. Wstępne szkolenie modelu	212
Krok 14. Zapisywanie przeszkolonego modelu (+ tokenizer + konfiguracja) na dysku	212
Krok 15. Modelowanie języka za pomocą potoku FillMaskPipeline	213
Wstępne szkolenie modelu obsługi klienta generatywnej sztucznej inteligencji na danych pochodzących z serwisu X	215
Krok 1. Pobieranie zbioru danych	216
Krok 2. Instalowanie bibliotek Hugging Face: transformers i datasets	216
Krok 3. Ładowanie i filtrowanie danych	216
Krok 4. Sprawdzanie ograniczeń zasobów: układ GPU i CUDA	218
Krok 5. Definiowanie konfiguracji modelu	218
Krok 6. Tworzenie i przetwarzanie zbioru danych	219
Krok 7. Inicjalizowanie obiektu trenera	220
Krok 8. Wstępne szkolenie modelu	221
Krok 9. Zapisywanie modelu	221
Krok 10. Interfejs użytkownika do czatu z agentem generatywnej AI	222
Dalsze szkolenie wstępne	224
Ograniczenia	224
Następne kroki	224
Podsumowanie	225

Pytania	226
Odnośniki	226
Lektura uzupełniająca	227
ROZDZIAŁ 7.	
ChatGPT — rewolucja w generatywnej sztucznej inteligencji	228
Model GPT jako technologia ogólnego przeznaczenia	229
Udoskonalenia	230
Rozpowszechnianie	231
Wszechobecność	232
Architektura modeli transformerów GPT firmy OpenAI	234
Rozwój modeli transformerów o miliardach parametrów	235
Coraz większe rozmiary modeli transformerów	235
Rozmiar kontekstu i maksymalna długość ścieżki	237
Od dostrajania do modeli zero-shot	238
Stos warstw dekodera	240
Modele GPT	241
Modele OpenAI w roli asystentów	244
ChatGPT udostępnia kod źródłowy	244
Asystent tworzenia kodu GitHub Copilot	245
Przykłady promptów ogólnego przeznaczenia	247
Rozpoczęcie pracy z ChatGPT — GPT-4 w roli asystenta	249
Rozpoczęcie pracy z API modelu GPT-4	254
Uruchomienie pierwszego zadania NLP z użyciem modelu GPT-4	254
Uruchamianie wielu zadań NLP	257
Wykorzystanie techniki RAG z GPT-4	258
Instalacja	258
Odzyskiwanie informacji z dokumentów	259
Zastosowanie techniki RAG	260
Podsumowanie	263
Pytania	264
Odnośniki	264

Lektura uzupełniająca	265
ROZDZIAŁ 8.	
Dostrajanie modeli GPT OpenAI	266
Zarządzanie ryzykiem	267
Dostrajanie modelu GPT do wykonywania (generatywnego) zadania uzupełniania	268
1. Przygotowywanie zbioru danych	270
1.1. Przygotowywanie danych w formacie JSON	270
1.2. Konwertowanie danych do formatu JSONL	272
2. Dostrajanie oryginalnego modelu	275
3. Uruchamianie dostrojonego modelu GPT	277
4. Zarządzanie zadaniami dostrajania i dostrojonymi modelami	280
Przed zakończeniem	281
Podsumowanie	283
Pytania	283
Odnośniki	284
Lektura uzupełniająca	284
ROZDZIAŁ 9.	
Rozbijanie czarnej skrzynki za pomocą narzędzi do interpretacji działania transformerów	285
Wizualizacja działania transformera z użyciem BertViz	287
Uruchamianie BertViz	287
Interpretacja działania transformerów Hugging Face za pomocą narzędzia SHAP	299
Podstawowe informacje o SHAP	299
Wyjaśnienie wyników transformerów Hugging Face z użyciem SHAP	302
Wizualizacja transformera poprzez uczenie słownikowe	304
Współczynniki transformera	304
Wprowadzenie do LIME	306
Interfejs wizualizacji	307
Inne narzędzia interpretacji mechanizmów AI	309

LIT	309
Modele LLM OpenAI wyjaśniają działanie neuronów w transformerach	313
Ograniczenia i kontrola ze strony człowieka	316
Podsumowanie	317
Pytania	317
Odnośniki	318
Lektura uzupełniająca	318
ROZDZIAŁ 10.	
Badanie roli tokenizerów w kształtowaniu modeli transformerów ...	319
Dopasowywanie zbiorów danych i tokenizerów	320
Najlepsze praktyki	321
Tokenizacja Word2Vec	325
Badanie tokenizerów zdań i tokenizerów WordPiece w celu zrozumienia wydajności tokenizerów podwyrazów w kontekście ich wykorzystania przez transformery	334
Tokenizerzy wyrazów i zdań	335
Tokenizerzy oparte na podwyrazach	339
Badanie tokenizerów w kodzie	343
Podsumowanie	348
Pytania	349
Odnośniki	349
Lektura uzupełniająca	349
ROZDZIAŁ 11.	
Wykorzystanie osadzeń LLM jako alternatywy dla precyzyjnego dostrajania	350
Osadzenia LLM jako alternatywa dla precyzyjnego dostrajania	352
Od projektowania promptów do inżynierii promptów	352
Podstawy osadzania tekstu za pomocą NLTK i Gensim	353
Instalowanie bibliotek	353
1. Odczytywanie pliku tekstowego	353

2. Tokenizacja tekstu z użyciem tokenizera Punkt	354
3. Osadzanie tekstu za pomocą Gensim i Word2Vec	356
4. Opis modelu	357
5. Dostęp do słowa i wektora słów	359
6. Analiza przestrzeni wektorowej Gensim	360
7. TensorFlow Projector	363
Implementacja systemów pytań i odpowiedzi z użyciem technik opartych na osadzeniach	367
1. Instalowanie bibliotek i wybór modeli	367
2. Implementacja modelu osadzeń i modelu GPT	368
3. Przygotowywanie danych do wyszukiwania	372
4. Wyszukiwanie	373
5. Zadawanie pytania	375
Uczenie transferowe z użyciem osadzeń Ada	379
1. Zbiór danych Amazon Fine Food Reviews	379
2. Obliczanie osadzeń Ada i zapisywanie ich w celu ponownego wykorzystania w przyszłości	381
3. Klasteryzacja	382
4. Próbkę tekstu w klastrach i nazwy klastrów	384
Podsumowanie	386
Pytania	387
Odnośniki	387
Lektura uzupełniająca	387
ROZDZIAŁ 12.	
Oznaczanie ról semantycznych bez analizy składniowej z wykorzystaniem modelu GPT-4 i ChatGPT	388
Rozpoczynanie pracy z technikami SRL	390
Wprowadzenie do świata AI bez składni	391
Definicja SRL	392
Wizualizacja SRL	393
Eksperymenty SRL z ChatGPT z modelem GPT-4	393

Prosty przykład	394
Trudny przykład	398
Kwestionowanie zakresu SRL	399
Wyzwania związane z analizą orzeczeń	400
Ponowna definicja SRL	401
Od technik SRL specyficznych dla zadania do emergencji z wykorzystaniem ChatGPT	403
1. Instalowanie OpenAI	404
2. Tworzenie funkcji dialogu z GPT-4	404
3. Uruchamianie żądań SRL	405
Podsumowanie	412
Pytania	413
Odnosińki	414
Lektura uzupełniająca	414
ROZDZIAŁ 13.	
Zadania generowania streszczeń z użyciem modeli T5 i ChatGPT	415
Projektowanie uniwersalnego modelu tekst – tekst	417
Powstanie modeli transformerów tekst – tekst	418
Prefiks zamiast formatów specyficznych dla zadań	419
Model T5	421
Tworzenie streszczeń tekstu z użyciem modelu T5	423
Hugging Face	423
Inicjalizowanie modelu transformera T5	425
Tworzenie streszczeń dokumentów z użyciem modelu T5	431
Od transformera tekst – tekst do prognoz nowych słów z użyciem systemu ChatGPT firmy OpenAI	438
Porównanie metod tworzenia streszczeń modelu T5 i systemu ChatGPT	438
Tworzenie streszczeń z użyciem ChatGPT	439
Podsumowanie	444
Pytania	444

Odknośniki	445
Lektura uzupełniajăca	445
ROZDZIAŁ 14.	
NajnowoczeŃniejsze modele LLM Vertex AI i PaLM 2	446
Architektura	448
Pathways	448
PaLM	451
PaLM 2	452
Asystenty AI	454
Gemini	456
Google Workspace	457
Google Colab Copilot	460
Interfejs Vertex AI modelu PaLM 2	462
API PaLM 2 Vertex AI	468
Odpowiadanie na pytania	469
Zadanie typu pytanie – odpowiedź	471
Podsumowanie dialogu	472
Analiza tonu	473
Zadania wielokrotnego wyboru	475
Kod	477
Dostrajanie	482
Utworzenie kontenera	483
Dostrajanie modelu	484
Podsumowanie	486
Pytania	487
Odknośniki	487
Lektura uzupełniajăca	487
ROZDZIAŁ 15.	
Pilnowanie gigantów, czyli łagodzenie zagrożeń zwiăzanych z uŹyciem modeli LLM	488
Powstanie funkcjonalnej sztucznej inteligencji ogólnej (AGI)	490

Ograniczenia instalacji najnowocześniejszych platform	492
Auto-BIG-bench	495
WandB	501
Kiedy agenty AI zaczną się replikować?	503
Zarządzanie zagrożeniami	505
Halucynacje i zapamiętywanie	506
Ryzykowne zachowania emergentne	511
Deinformacja	513
Wywieranie wpływu na opinię publiczną	514
Treści szkodliwe	516
Prywatność	518
Cyberbezpieczeństwo	519
Narzędzia do łagodzenia zagrożeń z RLHF i RAG	520
1. Moderowanie wejścia i wyjścia za pomocą transformerów i bazy reguł	521
2. Budowanie bazy wiedzy dla systemu ChatGPT i modelu GPT-4	525
3. Parsowanie żądań użytkownika i korzystanie z bazy wiedzy	527
4. Generowanie zawartości ChatGPT z funkcją obsługi dialogu	528
Podsumowanie	530
Pytania	531
Odnośniki	531
Lektura uzupełniająca	532
ROZDZIAŁ 16.	
Nie tylko tekst — transformery wizyjne u progu rewolucyjnej sztucznej inteligencji	533
Od modeli niezależnych od zadań do multimodalnych transformerów wizyjnych	534
Transformery wizyjne (ViT)	536
Podstawowa architektura ViT	536
Transformery wizyjne w kodzie	539
CLIP	549

Podstawowa architektura modelu CLIP	549
CLIP w kodzie	550
DALL-E 2 i DALL-E 3	553
Podstawowa architektura DALL-E	554
Wprowadzenie w tematykę API modeli DALL-E 2 i DALL-E 3	555
GPT-4V, DALL-E 3 i rozbieżne skojarzenia semantyczne	561
Definicja rozbieżnego skojarzenia semantycznego	561
Tworzenie obrazu z użyciem systemu ChatGPT Plus z DALL-E	563
Wykorzystanie API modelu GPT-4V i eksperymenty z zadaniami DAT ...	565
Podsumowanie	570
Pytania	571
Odnosińki	571
Lektura uzupełniająca	572
ROZDZIAŁ 17.	
Przekraczanie granic między obrazem a tekstem	
z użyciem modelu Stable Diffusion	573
Przekraczanie granic generowania obrazu	574
Część I. Zamiana tekstu na obraz z użyciem modelu Stable Diffusion	576
1. Osadzanie tekstu za pomocą kodera transformera	577
2. Tworzenie losowych obrazów z szumami	578
3. Próbkowanie w dół modelu Stable Diffusion	579
4. Próbkowanie w górę na poziomie dekodera	581
5. Wynikowy obraz	582
Uruchamianie implementacji Keras modelu Stable Diffusion	582
Część II. Zamiana tekstu na obraz za pomocą API Stable Diffusion	584
Wykorzystanie modelu Stable Diffusion generatywnej sztucznej	
inteligencji do wykonania zadania z zakresu	
skojarzeń rozbieżnych (DAT)	586
Część III. Zamiana tekstu na wideo	588
Zamiana tekstu na wideo z użyciem modeli animacji Stability AI	588
Zamiana tekstu na wideo z użyciem odmiany modelu CLIP	

firmy OpenAI	590
Zamiana wideo na tekst z użyciem modelu TimeSformer	592
Przygotowywanie klatek wideo	593
Wykorzystanie modelu TimeSformer do tworzenia prognoz na podstawie klatek wideo	595
Podsumowanie	596
Pytania	597
Odnośniki	597
Lektura uzupełniająca	597
ROZDZIAŁ 18.	
AutoTrain na platformie Hugging Face — szkolenie modeli wizyjnych bez kodowania	598
Cel i zakres tego rozdziału	600
Pierwsze kroki	601
Przesyłanie zestawu danych	602
Bez kodowania?	605
Szkolenie modeli za pomocą mechanizmu AutoTrain	606
Wdrażanie modelu	607
Uruchamianie modeli w celu wnioskowania	609
Pobieranie obrazów walidacyjnych	609
Wnioskowanie: klasyfikacja obrazów	611
Eksperymenty walidacyjne na przeszkolonych modelach	613
Wypróbowywanie skuteczności najlepszego modelu ViT dla korpusu obrazów	626
Podsumowanie	627
Pytania	628
Odnośniki	629
Lektura uzupełniająca	629
ROZDZIAŁ 19.	
Na drodze do funkcjonalnej ogólnej AI z systemem HuggingGPT i jego odpowiednikami	630

Definicja systemu F-AGI	632
Instalowanie i importowanie bibliotek	635
Zbiór walidacyjny	635
Poziom 1 — łatwy obraz	636
Poziom 2 — trudny obraz	636
Poziom 3 — bardzo trudny obraz	637
HuggingGPT	638
Poziom 1 — łatwy	640
Poziom 2 — trudny	642
Poziom 3 — bardzo trudny	645
CustomGPT	649
Google Cloud Vision	651
Łączenie modeli: Google Cloud Vision z ChatGPT	656
Łączenie modeli z użyciem systemu Runway Gen-2	658
Midjourney: wyobraź sobie okręt płynący w przestrzeni galaktycznej	659
Gen-2: niech ten statek pływa po morzu	660
Podsumowanie	661
Pytania	662
Odnośniki	663
Lektura uzupełniająca	663
ROZDZIAŁ 20.	
Nie tylko prompty projektowane przez człowieka — generatywne kreowanie pomysłów	664
Część I. Definicja generatywnego kreowania pomysłów	666
Zautomatyzowana architektura kreowania pomysłów	666
Zakres i ograniczenia	668
Część II. Automatyzacja projektowania promptów na potrzeby generatywnego projektowania obrazów	668
Prezentacja HTML systemu opartego na ChatGPT z modelem GPT-4	669
Llama 2	674

Wykorzystanie modelu Llama 2 z modelem Hugging Face	675
Midjourney	680
Microsoft Designer	687
Część III. Zautomatyzowane generatywne kreowanie pomysłów	
z użyciem modelu Stable Diffusion	690
1. Brak promptu, automatyczne instrukcje dla modelu GPT-4	691
2. Generowanie promptu przez generatywną sztuczną inteligencję	
z użyciem ChatGPT z modelem GPT-4	694
3. i 4. Generowanie obrazów przez generatywną sztuczną	
inteligencję z użyciem modelu Stable Diffusion i ich wyświetlanie	696
Przyszłość należy do Ciebie!	698
Przyszłość programistów dzięki technikom VR-AI	698
Podsumowanie	703
Pytania	704
Odnośniki	704
Lektura uzupełniająca	705
Dodatek. Odpowiedzi na pytania	706
Skorowidz	728