
Spis treści

| | |
|--|-----------|
| Wprowadzenie | 11 |
| 1. Podstawy generatywnej sztucznej inteligencji, przypadki użycia i cykl życia projektu | 15 |
| Przykłady użycia i zadania | 15 |
| Modele podstawowe i centra modeli | 18 |
| Cykl życia projektu generatywnej sztucznej inteligencji | 18 |
| Generatywna sztuczna inteligencja w chmurze AWS | 21 |
| Dlaczego chmura AWS? | 23 |
| Tworzenie aplikacji | |
| opartych na generatywnej sztucznej inteligencji w chmurze AWS | 24 |
| Podsumowanie | 26 |
| 2. Inżynieria monitu i uczenie kontekstowe | 27 |
| Monity i uzupełnienia | 27 |
| Tokeny | 28 |
| Inżynieria monitu | 28 |
| Struktura monitu | 29 |
| Instrukcja | 29 |
| Kontekst | 30 |
| Uczenie kontekstowe na kilku przykładach | 31 |
| Uczenie bez przykładów | 32 |
| Uczenie na jednym przykładzie | 32 |
| Uczenie na kilku przykładach | 32 |
| Błędne uczenie kontekstowe | 33 |
| Dobre praktyki uczenia kontekstowego | 34 |
| Dobre praktyki inżynierii monitu | 34 |
| Parametry wnioskowania | 38 |
| Podsumowanie | 42 |

| | |
|--|-----------|
| 3. Podstawowe duże modele językowe | 43 |
| Podstawowe duże modele językowe | 43 |
| Tokenizery | 45 |
| Wektory osadzeń | 46 |
| Architektura transformera | 47 |
| Dane wejściowe i okno kontekstowe | 48 |
| Osadzenia | 48 |
| Koder | 48 |
| Warstwy samouwagi | 49 |
| Dekoder | 50 |
| Funkcja softmax | 50 |
| Rodzaje modeli podstawowych opartych na transformerach | 52 |
| Zbiory danych do wstępnego trenowania modeli | 54 |
| Reguły skalowania | 55 |
| Modele optymalne obliczeniowo | 56 |
| Podsumowanie | 57 |
| 4. Optymalizacja pamięci i obliczeń | 59 |
| Wyzwania pamięciowe | 59 |
| Typy i precyzja danych | 61 |
| Kwantyzacja | 63 |
| Typ fp16 | 63 |
| Typ bfloat16 | 65 |
| Typ fp8 | 67 |
| Typ int8 | 67 |
| Optymalizacja warstw samouwagi | 69 |
| FlashAttention | 69 |
| Grouped-Query Attention | 70 |
| Rozproszone przetwarzanie danych | 71 |
| Algorytm DDP | 71 |
| Algorytm FSDP | 71 |
| Porównanie wydajności algorytmów FSDP i DDP | 74 |
| Rozproszone przetwarzanie danych w chmurze AWS | 75 |
| Algorytm FSDP w klastrze Amazon SageMaker | 76 |
| Pakiet AWS Neuron SDK i akcelerator AWS Trainium | 78 |
| Podsumowanie | 78 |
| 5. Dostrajanie i ocenianie modelu | 79 |
| Dostrajanie za pomocą instrukcji | 79 |
| Llama 2-Chat | 80 |
| Falcon-Chat | 80 |
| FLAN-T5 | 80 |

| | |
|--|------------|
| Zbiór instrukcji | 80 |
| Zbiór różnorodnych instrukcji | 80 |
| FLAN — przykładowy zbiór różnorodnych instrukcji | 81 |
| Szablon monitu | 83 |
| Konwersja niestandardowego zbioru danych w zbiór instrukcji | 83 |
| Dostrajanie modelu za pomocą niestandardowych instrukcji | 85 |
| Amazon SageMaker Studio | 86 |
| Amazon SageMaker JumpStart | 87 |
| Klasa Amazon SageMaker Estimator dla centrum Hugging Face | 87 |
| Ocenianie modelu | 89 |
| Wskaźniki skuteczności | 89 |
| Testy porównawcze i zbiory danych | 90 |
| Podsumowanie | 91 |
| 6. Dostrajanie PEFT | 93 |
| Dostrajanie pełne i PEFT | 93 |
| LoRA i QLoRA | 95 |
| Podstawy techniki LoRA | 96 |
| Rząd macierzy | 97 |
| Docelowe moduły i warstwy | 97 |
| Implementacja techniki LoRA | 98 |
| Scalanie adaptera LoRA z oryginalnym modelem | 99 |
| Osobne adaptory LoRA | 100 |
| Skuteczność dostrajania pełnego i LoRA | 100 |
| QLoRA | 101 |
| Zmiękczenie i dostrajanie monitu | 102 |
| Podsumowanie | 104 |
| 7. Metoda RLHF | 107 |
| Ludzkie wartości: przydatność, uczciwość, nieszkodliwość | 107 |
| Podstawy uczenia przez wzmacnianie | 108 |
| Niestandardowy system nagradzania | 110 |
| Gromadzenie danych treningowych z zaangażowaniem człowieka | 110 |
| Przykładowe instrukcje dla adnotatorów | 110 |
| Gromadzenie adnotacji | |
| z wykorzystaniem usługi Amazon SageMaker Ground Truth | 111 |
| Przygotowanie danych do wytrenowania systemu nagradzania | 113 |
| Trening systemu nagradzania | 115 |
| System nagradzania — detektor toksyczności firmy Meta | 116 |
| Dostrajanie modelu z wykorzystaniem techniki RLHF | 118 |
| Zastosowanie systemu nagradzania i techniki RLHF | 118 |
| Algorytm PPO | 119 |
| Dostrajanie modelu przy użyciu techniki RLHF i algorytmu PPO | 120 |

| | |
|--|------------|
| Zapobieganie hakowaniu nagród | 121 |
| Zastosowanie dostrajania PEFT i techniki RLHF | 123 |
| Ocenianie modelu dostrojonego z użyciem techniki RLHF | 123 |
| Ocena jakościowa | 124 |
| Ocena ilościowa | 124 |
| Załadowanie systemu oceniania | 125 |
| Definicja funkcji zwracającej zagregowaną ocenę | 125 |
| Porównanie ocen przed dostrojeniem i po nim | 126 |
| Podsumowanie | 127 |
| 8. Optymalizacja wdrożenia modelu | 129 |
| Optymalizacja modelu pod kątem wnioskowania | 129 |
| Przycinanie modelu | 130 |
| Kwantyzacje PTQ i GPTQ | 131 |
| Destylacja | 133 |
| Kontener LMI | 135 |
| AWS Inferentia: specjalny sprzęt do wnioskowania | 136 |
| Strategie aktualizowania i wdrażania modeli | 138 |
| Testy A/B | 138 |
| Wdrożenie równoległe | 139 |
| Wskaźniki i monitoring | 141 |
| Autoskalowanie | 141 |
| Zasady autoskalowania | 142 |
| Definiowanie zasady autoskalowania | 142 |
| Podsumowanie | 143 |
| 9. Aplikacje wnioskujące kontekstowo w oparciu o technikę RAG i agentów | 145 |
| Ograniczenia modeli LLM | 146 |
| Halucynacje | 146 |
| Odcięcie wiedzy | 147 |
| Generowanie wspomagane pobieraniem | 147 |
| Zewnętrzne źródła wiedzy | 148 |
| Proces RAG | 149 |
| Załadowanie dokumentów | 150 |
| Fragmentowanie dokumentów | 151 |
| Pobieranie dokumentów i ponowny ranking wyników | 151 |
| Rozszerzenie monitu | 152 |
| Koordynacja i implementacja techniki RAG | 153 |
| Ładowanie i fragmentowanie dokumentów | 154 |
| Magazyn wektorów osadzeń i pobieranie danych | 155 |
| Łańcuch pobrań | 158 |
| Ponowny ranking z wykorzystaniem algorytmu MMR | 161 |

| | |
|---|------------|
| Agent | 162 |
| Platforma ReAct | 163 |
| Platforma PAL | 165 |
| Aplikacje oparte na generatywnej sztucznej inteligencji | 168 |
| FMOps — utrzymanie cyklu życia projektu aplikacji | |
| opartej na generatywnej sztucznej inteligencji | 172 |
| Eksperymentowanie | 173 |
| Programowanie | 175 |
| Wdrożenie w środowisku produkcyjnym | 176 |
| Podsumowanie | 177 |
| 10. Multimodalne modele podstawowe | 179 |
| Zastosowania multimodalnych modeli generatywnej sztucznej inteligencji | 180 |
| Dobre praktyki inżynierii multimodalnego monitu | 180 |
| Generowanie i udoskonalanie obrazów | 181 |
| Generowanie obrazów | 181 |
| Edycja i udoskonalanie obrazów | 182 |
| Wrysowanie, rozrysowanie i podrysowanie obrazu | 187 |
| Wrysowanie obrazu | 187 |
| Rozrysowanie obrazu | 188 |
| Podrysowanie obrazu | 189 |
| Podpisywanie obrazów, moderowanie treści i odpowiadanie na wizualne pytania | 191 |
| Podpisywanie obrazów | 192 |
| Moderowanie treści | 192 |
| Odpowiadanie na wizualne pytania | 192 |
| Ocena modelu | 196 |
| Generatywna konwersja tekstu na obraz | 197 |
| Dyfuzja w przód | 199 |
| Rozumowanie niewerbalne | 199 |
| Podstawy algorytmu dyfuzyjnego | 201 |
| Dyfuzja w przód | 201 |
| Dyfuzja wstecz | 201 |
| Sieć U-Net | 202 |
| Model Stable Diffusion 2 | 203 |
| Koder tekstu | 205 |
| Sieć U-Net i proces dyfuzji | 206 |
| Kondycjonowanie tekstu | 207 |
| Uwaga krzyżowa | 207 |
| Harmonogram | 207 |
| Dekoder obrazu | 208 |

| | |
|--|------------|
| Model Stable Diffusion XL | 208 |
| Sieć U-Net i uwaga krzyżowa | 208 |
| Rafinator | 208 |
| Kondycjonowanie | 209 |
| Podsumowanie | 210 |
| 11. Sterowanie procesem generowania obrazów i dostrajanie modelu Stable Diffusion | 213 |
| ControlNet | 213 |
| Dostrajanie modelu | 218 |
| DreamBooth | 218 |
| Metody DreamBooth, PEFT i LoRA | 221 |
| Inwersja tekstu | 222 |
| Dostosowywanie modelu do ludzkich wartości przy użyciu techniki RLHF | 225 |
| Podsumowanie | 227 |
| 12. Amazon Bedrock — usługa zarządzana dla generatywnej sztucznej inteligencji | 229 |
| Modele podstawowe w usłudze Amazon Bedrock | 229 |
| Modele Amazon Titan | 230 |
| Modele Stability AI Stable Diffusion | 230 |
| Interfejs API usługi Amazon Bedrock do wnioskowania | 230 |
| Modele LLM | 232 |
| Generowanie kodu SQL | 232 |
| Streszczanie tekstu | 233 |
| Osadzenia | 234 |
| Dostrajanie modeli | 237 |
| Agenci | 239 |
| Modele multimodalne | 242 |
| Tworzenie obrazu z tekstu | 242 |
| Tworzenie obrazów z obrazów | 243 |
| Prywatność danych i bezpieczeństwo sieci | 245 |
| Zarządzanie i monitorowanie | 246 |
| Podsumowanie | 246 |