

Vorwort	13
Einleitung	15
Teil I: Einführung in Large Language Models	23
1 Überblick über Large Language Models	25
Was sind Large Language Models?	26
Definition von LLMs	28
Hauptmerkmale von LLMs	30
Wie LLMs funktionieren	33
Gängige moderne LLMs	42
BERT	42
GPT-3 und ChatGPT	43
T5	44
Domänenspezifische LLMs	45
Anwendungen von LLMs	46
Klassische NLP-Aufgaben	46
Freitexterzeugung	49
Informationsabruf/neuronale semantische Suche	50
Chatbots	51
Zusammenfassung	52
2 Semantische Suche mit LLMs	53
Die Aufgabe	54
Asymmetrische semantische Suche	55
Die Lösung im Überblick	56

Die Komponenten	57
Engines für Text-Embeddings	58
Chunking von Dokumenten	62
Vektordatenbanken	68
Pinecone	68
Open-Source-Alternativen	68
Neueinstufen der abgerufenen Ergebnisse	69
API	70
Alles zusammen	71
Performance	72
Die Kosten von Closed-Source-Komponenten	75
Zusammenfassung	75
3 Erstes Prompt Engineering und ein Chatbot mit ChatGPT	77
Prompt Engineering	77
Ausrichtung in Sprachmodellen	78
Einfach fragen	79
Few-Shot-Learning	81
Strukturierung der Ausgabe	82
Personas fordern auf	83
Mit Prompts modellübergreifend arbeiten	85
ChatGPT	85
Cohere	86
Open-Source-Prompt-Engineering	87
Einen Frage-Antwort-Bot mit ChatGPT aufbauen	89
Zusammenfassung	94
Teil II: Das Beste aus LLMs herausholen	97
4 LLMs mit individuellem Feintuning optimieren	99
Transfer Learning und Feintuning: die Grundlagen	100
Der Feintuning-Prozess im Detail	101
Vortrainierte Closed-Source-Modelle als Grundlage	103
Die OpenAI-API für das Feintuning	104
Die GPT-3-API für das Feintuning	104
Fallstudie 1: Stimmungsklassifizierung von Amazon-Rezensionen	105
Richtlinien und bewährte Methoden für Daten	105
Individuelle Beispiele mit der OpenAI-CLI vorbereiten	106
Die OpenAI-CLI einrichten	110
Hyperparameter auswählen und optimieren	110

Unser erstes feingetuntes LLM	111
Feingetunte Modelle mit quantitativen Metriken bewerten	111
Qualitative Bewertungstechniken	114
Feingetunte GPT-3-Modelle in Anwendungen integrieren	116
Fallstudie 2: Klassifizierung der Kategorien von	
Amazon-Rezensionen	116
Zusammenfassung	117
5 Fortgeschrittenes Prompt Engineering	119
Prompt-Injection-Angriffe	119
Eingaben und Ausgaben validieren	121
Beispiel: Validierungspipelines mit NLI aufbauen	122
Prompts im Stapel verarbeiten	125
Prompts verketteten	126
Verkettung als Schutz gegen Prompt Injection	129
Verkettung, um Prompt Stuffing zu verhindern	130
Beispiel: Sicherheit durch Verkettung multimodaler LLMs	132
Prompting mit Gedankenkette	134
Beispiel: Grundlegende Arithmetik	134
Noch einmal: Few-Shot-Learning	136
Beispiel: Grundscharithmetik mit LLMs	136
Testen und iterative Entwicklung von Prompts	146
Zusammenfassung	147
6 Embeddings und Modellarchitekturen anpassen	149
Fallstudie: Ein Empfehlungssystem aufbauen	150
Das Problem und die Daten einrichten	150
Das Problem der Empfehlung definieren	151
Unser Empfehlungssystem im Überblick	154
Ein benutzerdefiniertes Beschreibungsfeld generieren,	
um Artikel zu vergleichen	157
Mit Basis-Embeddern eine Baseline einrichten	159
Die Feintuning-Daten vorbereiten	159
Open-Source-Embedder mithilfe von Sentence Transformers	
feintunen	163
Zusammenfassung der Ergebnisse	165
Zusammenfassung	168

Teil III: Fortgeschrittene LLM-Nutzung	169
7 Jenseits der Basismodelle: LLMs kombinieren	171
Fallstudie: Visuelles Frage-Antwort-System	171
Einführung in unsere Modelle: der Vision Transformer, GPT-2 und DistilBERT	172
Projektion und Fusion verborgener Zustände	175
Was ist Cross-Attention, und warum ist sie entscheidend?	176
Unser benutzerdefiniertes multimodales Modell	179
Unsere Daten: Visual QA	182
Die VQA-Trainingsschleife	183
Zusammenfassung der Ergebnisse	184
Fallstudie: Reinforcement Learning from Feedback	186
Unser Modell: FLAN-T5	189
Unser Belohnungsmodell: Sentiment und grammatische Korrektheit	189
Die Bibliothek Transformer Reinforcement Learning	191
Die RLF-Trainingsschleife	192
Zusammenfassung der Ergebnisse	195
Zusammenfassung	196
8 Feintuning fortgeschrittener Open-Source-LLMs	197
Beispiel: Multilabel-Klassifizierung mit BERT für Anime-Genres	198
Die Performance für die Multilabel-Genre-Vorhersage von Anime-Titeln mit dem Jaccard-Koeffizienten messen	198
Eine einfache Feintuning-Schleife	200
Allgemeine Tipps zum Feintuning von Open-Source-LLMs	201
Zusammenfassung der Ergebnisse	209
Beispiel: LaTeX-Generierung mit GPT-2	211
Prompt Engineering für Open-Source-Modelle	212
Zusammenfassung der Ergebnisse	214
SAWYER: Sinans Versuch, kluge und dennoch fesselnde Antworten zu geben	215
Schritt 1: Überwachtes Feintuning mit Anweisungen	217
Schritt 2: Training des Belohnungsmodells	219
Schritt 3: Reinforcement Learning mit (geschätzter) menschlicher Rückkopplung	223
Zusammenfassung der Ergebnisse	224
Die sich ständig verändernde Welt des Feintunings	228
Zusammenfassung	229

9 LLMs in die Produktion überführen	231
Closed-Source-LLMs in der Produktion bereitstellen	231
Kostenprognosen	231
API-Schlüsselverwaltung	232
Open-Source-LLMs in der Produktion bereitstellen	232
Ein Modell für Inferenz vorbereiten	232
Interoperabilität	233
Quantisierung	234
Beschneiden	234
Wissensdestillation	234
Fallstudie: Unsere Anime-Genre-Vorhersage destillieren	236
Kostenprognosen mit LLMs	243
Die Plattform Hugging Face	243
Zusammenfassung	247
Ihre Beiträge sind wichtig	248
Weitermachen!	248
 Teil IV: Anhänge	 249
Anhang A: LLM-FAQs	251
Anhang B: LLM-Glossar	257
Anhang C: Archetypen von LLM-Anwendungen	263
 Index	 267