

Auf einen Blick

1	Vorwort	15
2	Einführung in die generative KI	33
3	Vortrainierte Modelle	67
4	Large Language Models	91
5	Prompt Engineering	153
6	Vektordatenbanken	181
7	Retrieval-Augmented Generation	253
8	Agentensysteme	301
9	Deployment	391
10	Ausblick	427

Inhalt

Materialien zum Buch	13
----------------------------	----

1	Vorwort	15
----------	----------------	-----------

1.1	Zielsetzung des Buches	16
1.2	Zielgruppe	17
1.3	Was Sie schon wissen sollten	18
1.3.1	Verpflichtende Voraussetzungen	18
1.3.2	Optionale Voraussetzungen	18
1.4	Struktur des Buches	18
1.5	Wie man dieses Buch effektiv nutzt	22
1.6	Code zum Herunterladen und weitere Materialien	23
1.7	Systemeinrichtung	23
1.7.1	Installation von Python	24
1.7.2	Installation der IDE	24
1.7.3	Installation von git	26
1.7.4	Download des Kursmaterials	26
1.7.5	Einrichtung der lokalen Python-Umgebung	27
1.8	Danksagung	30
1.9	Konventionen in diesem Buch	30

2	Einführung in die generative KI	33
----------	--	-----------

2.1	Einführung in die künstliche Intelligenz	36
2.2	Die Säulen des Fortschritts in der generativen KI	40
2.2.1	Rechenleistung	40
2.2.2	Investitionen	42
2.2.3	Verbesserung der Algorithmen	42
2.3	Deep Learning	43
2.4	Schwache und allgemeine KI	46
2.5	Natural Language Processing (NLP)	49

- 2.5.1 NLP-Aufgaben 49
 - 2.5.2 Architekturen 53
- 2.6 Large Language Models (LLMs) 55**
 - 2.6.1 Training 55
 - 2.6.2 Daten 55
 - 2.6.3 Vortrainiertes Modell 56
 - 2.6.4 Instruktionsmodell (Instruction model) 56
 - 2.6.5 Sicherheitsmodell (Safety model) 56
 - 2.6.6 Evaluierung 56
- 2.7 Use-Cases 57**
 - 2.7.1 Kreatives Schreiben 57
 - 2.7.2 Zusammenfassung 57
 - 2.7.3 Chatbots 58
 - 2.7.4 Übersetzung 58
 - 2.7.5 Inhaltserzeugung 58
 - 2.7.6 Bildung 59
- 2.8 Die Grenzen von LLMs 59**
 - 2.8.1 Halluzinationen 59
 - 2.8.2 Verzerrungen und Vorurteile (Biases) 59
 - 2.8.3 Kontextfenster und Token-Beschränkung (Context Window, Token Length Constraints) 59
- 2.9 Large Multimodal Models (LMMs) 60**
 - 2.9.1 Anwendungen 61
 - 2.9.2 Herausforderungen 61
- 2.10 Generative KI-Anwendungen 62**
 - 2.10.1 Konsumenten 63
 - 2.10.2 Business 63
 - 2.10.3 Prosumer 64
- 2.11 Zusammenfassung 64**

3 Vortrainierte Modelle 67

- 3.1 Was sind vortrainierte Modelle? 69**
- 3.2 Hugging Face 69**
- 3.3 Modellauswahl 70**
- 3.4 Coding: Textzusammenfassung 71**

3.5	Übung: Übersetzung	73
3.6	Coding: Zero-Shot-Klassifikation	74
3.7	Coding: Füllmaske	78
3.8	Coding: Frage-Antwort Modelle	79
3.9	Coding: Erkennung bekannter Entitäten (Named Entity Recognition)	81
3.10	Coding: Text-zu-Bild	83
3.11	Übung: Text-zu-Audio	85
3.12	Abschlussprojekt: Kunden-Feedback analysieren	86
3.13	Zusammenfassung	89

4	Large Language Models	91
----------	------------------------------	----

4.1	Eine kurze Geschichte der Sprachmodelle	92
4.2	LLMs mithilfe von Python nutzen	93
4.2.1	Coding: OpenAI nutzen	94
4.2.2	Coding: Groq nutzen	97
4.2.3	Multimodale Modelle	101
4.2.4	Coding: Multimodale Modelle	101
4.2.5	Coding: LLMs lokal betreiben	104
4.3	Modellparameter	107
4.3.1	Die Modelltemperatur	108
4.3.2	Top-p und Top-k	110
4.3.3	Empfehlungen	111
4.4	Modellauswahl	111
4.4.1	Leistungsfähigkeit	112
4.4.2	Wissensstand	113
4.4.3	On-Premises- vs. Cloud-Hosting	113
4.4.4	Open-Source, Open-Weight und proprietäre Modelle	113
4.4.5	Preis	114
4.4.6	Kontextfenster	114
4.4.7	Latenz	114
4.5	Messages	115
4.5.1	User	115
4.5.2	System	115
4.5.3	Assistent	116

- 4.6 Prompt Templates** 116
 - 4.6.1 Coding: ChatPromptTemplates 116
 - 4.6.2 Coding: Verbesserung eines Prompts mit LangChain Hub 118
- 4.7 Chains** 120
 - 4.7.1 Coding: Eine einfache sequenzielle Chain 121
 - 4.7.2 Coding: Parallele Chains 123
 - 4.7.3 Coding: Router-Chain 125
 - 4.7.4 Coding: Chain mit Gedächtnis 131
- 4.8 LLM-Schutz und -Sicherheit** 135
 - 4.8.1 LLM Sicherheit 136
 - 4.8.2 LLM-Schutz 136
- 4.9 Modellverbesserungen** 143
- 4.10 Neue Trends** 144
 - 4.10.1 Reasoning-Modelle 145
 - 4.10.2 Small-Language-Modelle 146
 - 4.10.3 Test-Time Compute 148
- 4.11 Zusammenfassung** 151

5 Prompt Engineering 153

- 5.1 Prompting – die Grundlagen** 154
 - 5.1.1 Prompt und Prompt Templates 154
 - 5.1.2 Der Prompt-Engineering-Prozess 154
 - 5.1.3 Vom Prompt zum Modell-Ergebnis 155
 - 5.1.4 Prompt-Komponenten 156
 - 5.1.5 Grundprinzipien 157
- 5.2 Coding: Few-Shot Prompting** 163
- 5.3 Chain-of-Thought** 166
- 5.4 Zero-Shot Chain-of-Thought** 166
- 5.5 Coding: Self-Consistency Chain-of-Thought** 167
- 5.6 Coding: Prompt-Chaining** 171
- 5.7 Coding: Self-Feedback** 173
- 5.8 Zusammenfassung** 178

6	Vektordatenbanken	181
6.1	Einleitung	181
6.2	Der Data-Ingestion-Prozess	184
6.3	Dokumente importieren	185
6.3.1	Ein erster Überblick	185
6.3.2	Coding: Eine einzelne Textdatei laden	186
6.3.3	Coding: Laden mehrerer Textdateien	188
6.3.4	Übung: Laden mehrerer Wikipedia Artikel	189
6.3.5	Übung: Laden von Büchern von Project Gutenberg	191
6.4	Dokumente aufteilen	193
6.4.1	Coding: Chunking mit festen Größen	195
6.4.2	Coding: Strukturbasiertes Chunking	199
6.4.3	Coding: Semantisches Chunking	202
6.4.4	Coding: Benutzerdefiniertes Chunking	205
6.5	Einbettungen erstellen	209
6.5.1	Überblick	210
6.5.2	Coding: Wort-Einbettungen	211
6.5.3	Coding: Satzeinbettungen	219
6.5.4	Coding: Einbettungen mit LangChain erzeugen	222
6.6	Daten speichern	225
6.6.1	Auswahl einer Vektordatenbank	225
6.6.2	Coding: Lokale Vektordatenbank mit ChromaDB	226
6.6.3	Coding: Cloud-basierte Vektor-DB mit Pinecone	228
6.7	Daten abrufen	231
6.7.1	Berechnung der Ähnlichkeit	232
6.7.2	Coding: Rückgabe von Daten mittels ChromaDB	234
6.7.3	Coding: Rückgabe von Daten mittels Pinecone	236
6.8	Abschlussprojekt	238
6.8.1	Die Features der App	240
6.8.2	Datensatz	241
6.8.3	Vorbereitung der Vektordatenbank	241
6.8.4	Übung: Alle Genres aus der Vektordatenbank extrahieren	245
6.8.5	App-Entwicklung	246
6.9	Zusammenfassung	251

7	Retrieval-Augmented Generation	253
7.1	Einleitung	254
7.1.1	Retrieval-Prozess	256
7.1.2	Augmentierung	257
7.1.3	Generierung	257
7.2	Ein einfaches System zur Retrieval-Augmented Generation	258
7.2.1	Vorbereitung der Wissensbasis	258
7.2.2	Retrieval	260
7.2.3	Augmentierung	261
7.2.4	Generierung	262
7.2.5	Erstellung der RAG-Funktion	263
7.3	Fortgeschrittene Techniken	265
7.3.1	Fortgeschrittene Prä-Retrieval-Techniken	265
7.3.2	Fortgeschrittene Retrieval-Techniken	269
7.3.3	Fortgeschrittene Post-Retrieval-Techniken	286
7.4	Coding: Prompt-Caching	287
7.5	Evaluierung	293
7.5.1	Herausforderungen in der RAG-Evaluierung	293
7.5.2	Metriken	294
7.5.3	Coding: Metriken	296
7.6	Zusammenfassung	299
8	Agentensysteme	301
8.1	Einführung in KI-Agenten	302
8.2	Verfügbare Frameworks	304
8.3	Einfache Agentensysteme	306
8.3.1	Agentenbasiertes RAG	306
8.3.2	ReAct	310
8.4	Agenten-Framework: LangGraph	314
8.4.1	Einfacher Graph: Assistent	315
8.4.2	Router-Graph	320
8.4.3	Graph mit Tools	324

8.5	Agenten-Framework: AG2	330
8.5.1	Konversation zweier Agenten	331
8.5.2	Human-in-the-Loop	335
8.5.3	Agenten, die Tools benutzen	342
8.6	Agenten-Framework: CrewAI	346
8.6.1	Einleitung	346
8.6.2	Die erste Crew: eine Nachrichteanalyse-Crew	348
8.6.3	Übung: KI-Sicherheits-Crew	364
8.7	Agenten-Framework: OpenAI Agents	374
8.7.1	Erste Schritte mit einem einzelnen Agenten	374
8.7.2	Arbeit mit mehreren Agenten	375
8.7.3	Agent mit Such- und Dateiabruf-Funktionalität	377
8.8	Agenten-Framework: Pydantic AI	379
8.9	Überwachung von Agentensystemen	382
8.9.1	AgentOps	382
8.9.2	Logfire	386
8.10	Zusammenfassung	388

9	Deployment	391
----------	-------------------	-----

9.1	Die Anwendungsarchitektur	392
9.2	Die Deploymentstrategie	394
9.2.1	REST-API-Entwicklung	394
9.2.2	Deployment-Ziele	395
9.2.3	Coding: Lokales Deployment	397
9.3	Entwicklung einer eigenständigen Anwendung	403
9.4	Deployment auf Heroku	410
9.4.1	Erstellen einer neuen App	410
9.4.2	Download und Konfiguration der CLI	411
9.4.3	Eine app.py-Datei erzeugen	412
9.4.4	Procfile-Setup	414
9.4.5	Umgebungsvariablen	414
9.4.6	Die Python-Umgebung	415
9.4.7	Lokale Überprüfung des Ergebnisses	415

9.4.8	Los geht's: Das Deployment auf Heroku	416
9.4.9	Die App stoppen	417
9.5	Deployment auf Streamlit.io	419
9.5.1	Das GitHub-Repository anlegen	419
9.5.2	Eine neue App erstellen	420
9.6	Deployment auf Render	421
9.7	Zusammenfassung	424
 10 Ausblick		427
10.1	Fortschritte in der Modellarchitektur	427
10.2	Limitierungen und Probleme von LLMs	428
10.2.1	Halluzinationen	428
10.2.2	Vorurteile (Biases)	429
10.2.3	Falschinformationen	431
10.2.4	Geistiges Eigentum	432
10.2.5	Interpretierbarkeit und Transparenz	432
10.2.6	Jailbreaking LLMs	432
10.3	Regulatorische Entwicklungen	434
10.4	Künstliche allgemeine Intelligenz und künstliche Super-Intelligenz	434
10.5	KI-Fähigkeiten in der nahen Zukunft	435
10.6	Hilfreiche Ressourcen	438
10.7	Zusammenfassung	439
 Über den Autor		441
Index		443