

# Generative KI mit Python

RAG-Anwendungen und Agentensysteme mit  
Vektordatenbanken und LLMs

# DAS INHALTS- VERZEICHNIS

» Hier geht's  
direkt  
zum Buch

# Inhalt

Materialien zum Buch .....	13
<b>1 Vorwort</b> .....	<b>15</b>
<b>1.1 Zielsetzung des Buches</b> .....	<b>16</b>
<b>1.2 Zielgruppe</b> .....	<b>17</b>
<b>1.3 Was Sie schon wissen sollten</b> .....	<b>18</b>
1.3.1 Verpflichtende Voraussetzungen .....	18
1.3.2 Optionale Voraussetzungen .....	18
<b>1.4 Struktur des Buches</b> .....	<b>18</b>
<b>1.5 Wie man dieses Buch effektiv nutzt</b> .....	<b>22</b>
<b>1.6 Code zum Herunterladen und weitere Materialien</b> .....	<b>23</b>
<b>1.7 Systemeinrichtung</b> .....	<b>23</b>
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
<b>1.8 Danksagung</b> .....	<b>30</b>
<b>1.9 Konventionen in diesem Buch</b> .....	<b>30</b>
<b>2 Einführung in die generative KI</b> .....	<b>33</b>
<b>2.1 Einführung in die künstliche Intelligenz</b> .....	<b>36</b>
<b>2.2 Die Säulen des Fortschritts in der generativen KI</b> .....	<b>40</b>
2.2.1 Rechenleistung .....	40
2.2.2 Investitionen .....	42
2.2.3 Verbesserung der Algorithmen .....	42
<b>2.3 Deep Learning</b> .....	<b>43</b>
<b>2.4 Schwache und allgemeine KI</b> .....	<b>46</b>
<b>2.5 Natural Language Processing (NLP)</b> .....	<b>49</b>

- 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**

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

## **4 Large Language Models** 91

---

<b>4.1 Eine kurze Geschichte der Sprachmodelle</b> .....	92
<b>4.2 LLMs mithilfe von Python nutzen</b> .....	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
<b>4.3 Modellparameter</b> .....	107
4.3.1 Die Modelltemperatur .....	108
4.3.2 Top-p und Top-k .....	110
4.3.3 Empfehlungen .....	111
<b>4.4 Modellauswahl</b> .....	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
<b>4.5 Messages</b> .....	115
4.5.1 User .....	115
4.5.2 System .....	115
4.5.3 Assistent .....	116

<b>4.6</b>	<b>Prompt Templates</b> .....	116
4.6.1	Coding: ChatPromptTemplates .....	116
4.6.2	Coding: Verbesserung eines Prompts mit LangChain Hub .....	118
<b>4.7</b>	<b>Chains</b> .....	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
<b>4.8</b>	<b>LLM-Schutz und -Sicherheit</b> .....	135
4.8.1	LLM Sicherheit .....	136
4.8.2	LLM-Schutz .....	136
<b>4.9</b>	<b>Modellverbesserungen</b> .....	143
<b>4.10</b>	<b>Neue Trends</b> .....	144
4.10.1	Reasoning-Modelle .....	145
4.10.2	Small-Language-Modelle .....	146
4.10.3	Test-Time Compute .....	148
<b>4.11</b>	<b>Zusammenfassung</b> .....	151

## **5 Prompt Engineering** 153

---

<b>5.1</b>	<b>Prompting – die Grundlagen</b> .....	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
<b>5.2</b>	<b>Coding: Few-Shot Prompting</b> .....	163
<b>5.3</b>	<b>Chain-of-Thought</b> .....	166
<b>5.4</b>	<b>Zero-Shot Chain-of-Thought</b> .....	166
<b>5.5</b>	<b>Coding: Self-Consistency Chain-of-Thought</b> .....	167
<b>5.6</b>	<b>Coding: Prompt-Chaining</b> .....	171
<b>5.7</b>	<b>Coding: Self-Feedback</b> .....	173
<b>5.8</b>	<b>Zusammenfassung</b> .....	178

<b>6</b>	<b>Vektordatenbanken</b>	181
<b>6.1</b>	<b>Einleitung</b>	181
<b>6.2</b>	<b>Der Data-Ingestion-Prozess</b>	184
<b>6.3</b>	<b>Dokumente importieren</b>	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
<b>6.4</b>	<b>Dokumente aufteilen</b>	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
<b>6.5</b>	<b>Einbettungen erstellen</b>	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
<b>6.6</b>	<b>Daten speichern</b>	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
<b>6.7</b>	<b>Daten abrufen</b>	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
<b>6.8</b>	<b>Abschlussprojekt</b>	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
<b>6.9</b>	<b>Zusammenfassung</b>	251

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

<b>8.5</b>	<b>Agenten-Framework: AG2</b> .....	330
8.5.1	Konversation zweier Agenten .....	331
8.5.2	Human-in-the-Loop .....	335
8.5.3	Agenten, die Tools benutzen .....	342
<b>8.6</b>	<b>Agenten-Framework: CrewAI</b> .....	346
8.6.1	Einleitung .....	346
8.6.2	Die erste Crew: eine Nachrichteanalyse-Crew .....	348
8.6.3	Übung: KI-Sicherheits-Crew .....	364
<b>8.7</b>	<b>Agenten-Framework: OpenAI Agents</b> .....	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
<b>8.8</b>	<b>Agenten-Framework: Pydantic AI</b> .....	379
<b>8.9</b>	<b>Überwachung von Agentensystemen</b> .....	382
8.9.1	AgentOps .....	382
8.9.2	Logfire .....	386
<b>8.10</b>	<b>Zusammenfassung</b> .....	388
<b>9</b>	<b>Deployment</b> .....	391
<b>9.1</b>	<b>Die Anwendungsarchitektur</b> .....	392
<b>9.2</b>	<b>Die Deploymentstrategie</b> .....	394
9.2.1	REST-API-Entwicklung .....	394
9.2.2	Deployment-Ziele .....	395
9.2.3	Coding: Lokales Deployment .....	397
<b>9.3</b>	<b>Entwicklung einer eigenständigen Anwendung</b> .....	403
<b>9.4</b>	<b>Deployment auf Heroku</b> .....	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
<b>9.5</b>	<b>Deployment auf Streamlit.io .....</b>	<b>419</b>
9.5.1	Das GitHub-Repository anlegen .....	419
9.5.2	Eine neue App erstellen .....	420
<b>9.6</b>	<b>Deployment auf Render .....</b>	<b>421</b>
<b>9.7</b>	<b>Zusammenfassung .....</b>	<b>424</b>

## **10 Ausblick** 427

---

<b>10.1</b>	<b>Fortschritte in der Modellarchitektur .....</b>	<b>427</b>
<b>10.2</b>	<b>Limitierungen und Probleme von LLMs .....</b>	<b>428</b>
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
<b>10.3</b>	<b>Regulatorische Entwicklungen .....</b>	<b>434</b>
<b>10.4</b>	<b>Künstliche allgemeine Intelligenz und künstliche Super-Intelligenz .....</b>	<b>434</b>
<b>10.5</b>	<b>KI-Fähigkeiten in der nahen Zukunft .....</b>	<b>435</b>
<b>10.6</b>	<b>Hilfreiche Ressourcen .....</b>	<b>438</b>
<b>10.7</b>	<b>Zusammenfassung .....</b>	<b>439</b>

Über den Autor .....	441
Index .....	443