

# Contents

**Series Editor Foreword** xi

**Preface** xiii

**Acknowledgments** xvii

**About the Author** xix

## **I Getting Started with Foundations of AI, LLMs, and Experimentation** 1

### **1 An Introduction to AI, LLMs, and Agents** 3

Introduction 3

The Basics of Large Language Models 3

    What Is a Language Model—and Is It the “AI”  
    Covered in This Book? 4

    Parameter Count 6

    Context Windows 7

The Family Tree of LLM Tasks 10

Alignment 10

Prompt Engineering 12

    Prompt Ordering 12

    Chain of Thought 14

    Few-Shot Learning 14

    Prompt Chaining 16

Special LLM Features 17

    Inference Parameters 17

    Prompt Caching 19

    Structured Outputs 23

    Tool/Function Calling 24

LLM Workflows 25

AI Agents 25

    ReAct Agents 26

Conclusion 28

**2 First Steps with LLM Workflows 31**

Introduction 31

Case Study 1: Text-to-SQL Workflow 32

Exploratory Data Analysis 35

The First LLM Workflow 37

Step 1: Indexing Evidence into a Database 39

Step 2: Setting Up LangGraph 41

Step 3: Implementing Evidence Retrieval 42

Step 4: Writing the Prompt for SQL  
Generation 46

Step 5: Putting It All Together 48

Using the RAG Workflow 52

Conclusion 57

**3 AI Evaluation Plus Experimentation 59**

Introduction 59

Evaluating and Experimenting with LLMs 59

Case Study 1, Revisited: The Text-to-SQL Workflow 61

Evaluating SQL Generation: Free Text  
Response 61

Evaluating Evidence Retrieval 68

Evaluating Domain Difficulty 75

Case Study 2: A “Simple” Summary Prompt 77

Experiment: Prompt Chaining Summaries 80

Conclusion 83

**II Moving the Needle with AI Agents, Workflows, and  
Multimodality 85****4 First Steps with AI Agents and Multi-Agent  
Workloads 87**

Introduction 87

Case Study 3: From RAG to Agents 88

Defining Our Tools 90

Evaluating Our SQL Agent 95

Experiment: The Extended Mind Thesis and  
Agentic Memory 100

When Should You Use Workflows Versus Agents? 104

Case Study 4: A (Nearly) End-to-End SDR 105

Agent 1: Lead Generation 105

Agent 2: Lead Qualification 112

Agent 3: Lead Emailing 114

When to Use a Multi-Agent Versus a Single Agent 116

Evaluating Agents 118

LangSmith for Traceability 119

Conclusion 121

## **5 Enhancing Agents with Prompting, Workflows, and More Agents 123**

Introduction 123

Case Study 5: Agents Complying with Policies Plus Synthetic Data Generation 124

Creating a Test Set for Policy Compliance 124

Building Our Policy Bot Agent 127

BM25: Keeping It Old School 127

Prompt Engineering Agents 130

Evaluating Our Agents on Response Quality and Instructional Alignment 131

Case Study 6: Deep Research Plus Content Generation Agentic Workflows 133

Planning Components 133

Reflection Components 135

The Deep Research Agentic Workflow 136

Using Deep Research to Write a Custom Newsletter 140

Multi-Agent Architectures 141

Example: A Network-Based Multi-Agent Architecture 143

Case Study 4, Revisited: Adding a Supervisor Agent to Our SDR Team 148

Case Study 7: Agentic Tool Selection Performance 149

Investing in Tool Selection Accuracy, Precision, and Recall 152

Positional Bias in Tool Selection 156

Conclusion 157

## **6 Moving Beyond Natural Language: Multimodal and Coding AI 159**

Introduction 159

Introduction to Multimodal AI 159

Embed Modalities in the Same Vector Space 160

Map from One Mode to Another 164

Ground Modalities into a Primary Modality 165

Jointly Model Modes of Data 165

Handle Modalities Separately 168

Case Study 8: Image Retrieval Pipelines 168

Case Study 9: Visual Q/A with Moondream 174

Case Study 10: Coding Agent with Image Generation, File Use, and Moondream 176

Building a Coding Agent with Inception's Mercury Diffusion LLM 178

Giving Our Agent the Ability to **Generate** Images 184

Adding in Moondream Access 187

The Case for Any-to-Any Models 188

Conclusion 191

## **III Optimizing Workloads with Fine-Tuning, Frameworks, and Reasoning LLMs 193**

### **7 Reasoning LLMs and Computer Use 195**

Introduction 195

Seven Pillars of Intelligence 195

The Context Engineering Framework 197

Case Study 11: Benchmarking Reasoning Models 198

When Reasoning Helps 201

Comparing LLM Reasoning Efforts on HLE 205

Prompting Reasoning LLMs with MathQA 206

Reasoning Models for ReAct Agents 210

Case Study 12: Computer Use 212

Truly Multimodal Versus Grounded Computer Use 213

Benchmarking Computer Use with Reasoning Models 215

Building Computer Use with LangGraph 220

The Final Verdict on Reasoning LLMs 223

Conclusion 224

## **8 Fine-Tuning AI for Calibrated Performance 225**

Introduction 225

Case Study 13: Classification Versus Multiple Choice 227

Introducing the app\_reviews Dataset 229

LLM Calibration 231

Evaluating the Baseline LLMs on the Test Set 234

Fine-Tuning the LLM 235

Comparing Cost and Speed 238

Data Privacy 242

Balancing Accuracy, Cost, Speed, and Privacy 242

Calibration in Free-Text Responses 243

Case Study 14: Domain Adaptation 245

Chunking Policy Documents 248

Fine-Tuning a Qwen3 Reasoning LLM on Airbnb Policies 252

Evaluating the Domain-Adapted LLM 257

Conclusion 258

## **9 Optimizing AI Models for Production 261**

Introduction 261

Model Compression 261

Quantization 262

Distillation 266

Case Study 15: Speculative Decoding with Qwen 269

Case Study 16: Voice Bot—Need for Speed 272

Finding the Fastest STT and TTS Models 273

Case Study 17: Fine-Tuning Matryoshka Embeddings 277

Matryoshka Embeddings	279
Experimenting with Different Training Recipes	281
The Final Results	283
Case Study <i>N</i> + 1: What Comes Next?	284

**Index 287**