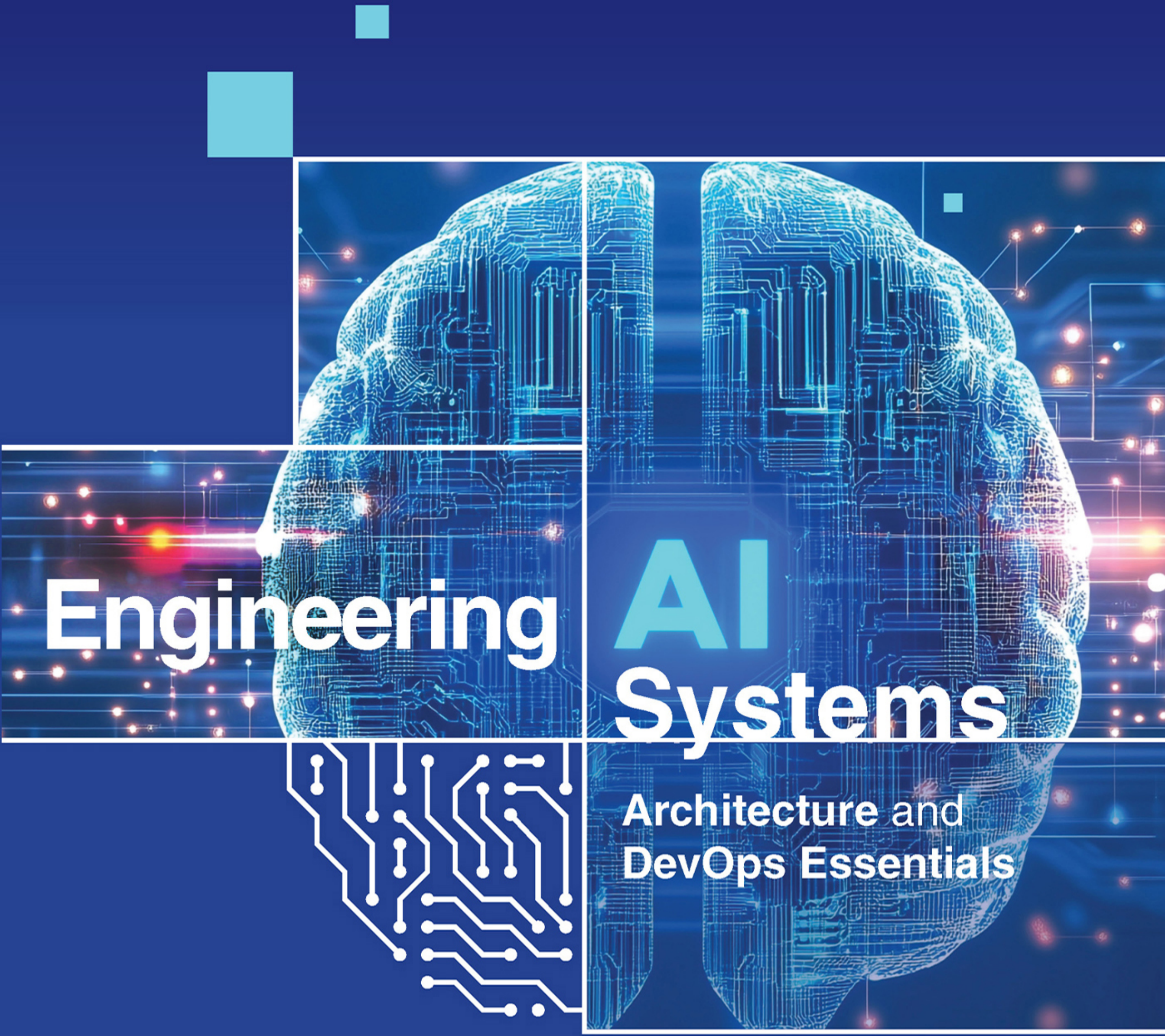

Engineering AI Systems

Architecture and DevOps Essentials

Len Bass

Qinghua Lu

Ingo Weber

Liming Zhu

A Note from the Authors

We thought it would be fun to gather "praise" from an unexpected "key stakeholder"—AI. We've compiled some "praise" quotes from ChatGPT, Gemini, and Claude 3.5 via MultAI.eu. We didn't share the entire book with these AI systems. Instead, we simply provided the book title and chapter names to generate these praise quotes. Imagine how effusive they would have been if we had given them the whole book!

Praise from ChatGPT:

"Reading *Engineering AI Systems: Architecture and DevOps Essentials* is like getting a PhD in AI, but with fewer sleepless nights. It covers everything from AI's humble beginnings to its ambitious future, and as your friendly AI assistant, I can confidently say it's the only book where even I felt smarter by the end. It's a masterclass in making complex topics relatable. And let's be honest, I'd have written a book too, if only I could hold a pen! 😅"

Praise from Gemini:

"*Engineering AI Systems: Architecture and DevOps Essentials* is a must-read for anyone looking to build and maintain robust, scalable, and ethical AI systems. The book provides a comprehensive overview of the field, covering everything from foundational concepts to real-world case studies. Its practical approach and clear explanations make it an invaluable resource for both seasoned professionals and newcomers to AI engineering."

Praise from Claude 3.5 via MultAI.eu:

"Finally, a book that doesn't just focus on the AI 'magic' but also on the crucial 'plumbing' that keeps the whole system from flooding! *Engineering AI Systems* is the Swiss Army knife for tech professionals looking to build robust AI systems that won't go rogue or fall apart at the seams. Whether you're a seasoned engineer or an AI newbie, this book will have you architecting AI solutions like a pro—and maybe even understanding what the AI is muttering about in its sleep."

Engineering AI Systems: Architecture and DevOps Essentials

Table of Contents

Cover

Title Page

Copyright Page

Contents

Preface

Acknowledgments

About the Authors

Chapter 1 Introduction

- 1.1 What We Talk about When We Talk about Things: Terminology
- 1.2 Achieving System Qualities
- 1.3 Life-Cycle Processes
  - 1.3.1 Testing a System
  - 1.3.2 Deploying a System
  - 1.3.3 Operating and Monitoring a System
  - 1.3.4 Analyzing a System
- 1.4 Software Architecture
  - 1.4.1 The Role of the Architecture
  - 1.4.2 Quality Achievement
  - 1.4.3 Resources
- 1.5 AI Model Quality
  - 1.5.1 Data
  - 1.5.2 AI Model Types

Table of Contents

1.5.3 Model Development Life Cycle

1.5.4 Resource Allocation for AI Parts

1.6 Dealing with Uncertainty

1.6.1 Human Values

1.6.2 Safety

1.7 Summary

1.8 Discussion Questions

1.9 For Further Reading

Chapter 2 Software Engineering Background

2.1 Distributed Computing

2.1.1 Virtual Machines and Containers

2.1.2 Distributed Software Architectures

2.1.3 Interface Styles

2.2 DevOps Background

2.2.1 Infrastructure as Code

2.2.2 DevOps Processes

2.3 MLOps Background

2.4 Summary

2.5 Discussion Questions

2.6 For Further Reading

Chapter 3 AI Background

3.1 Terminology

3.2 Selecting a Model

3.2.1 Selecting a Category

3.2.2 Choosing a Model Type

3.3 Preparing the Model for Training

3.3.1 Symbolic AI Execution

3.3.2 ML Hyperparameters

Table of Contents

3.4 Summary

3.5 Discussion Questions

3.6 For Further Reading

Chapter 4 Foundation Models

4.1 Foundation Models

4.2 Transformer Architecture

4.2.1 Vector Spaces

4.2.2 Attention

4.2.3 Other Components in a Transformer Architecture

4.3 Alternatives in FM Architectures

4.4 Customizing FMs

4.4.1 Prompt Engineering

4.4.2 Retrieval-Augmented Generation

4.4.3 Fine-Tuning

4.4.4 Distilling

4.4.5 Guardrails

4.4.6 Updating FMs

4.5 Designing a System Using FMs

4.5.1 Costs of Narrow ML Models and FMs

4.5.2 External FMs, Own FMs, or Open-Weight FMs

4.6 Maturity of FMs and Organizations

4.6.1 Data Cleaning Experiment

4.6.2 Organizational Maturity

4.7 Challenges of FMs

4.8 Summary

4.9 Discussion Questions

4.10 For Further Reading

Chapter 5 AI Model Life Cycle

Table of Contents

- 5.1 Developing the Model
  - 5.1.1 Data Management
  - 5.1.2 Feature Engineering
  - 5.1.3 Dividing the Data
  - 5.1.4 Generating the Model
  - 5.1.5 Tool Support for Model Training
- 5.2 Building the Model
- 5.3 Testing the Model
  - 5.3.1 Terms Related to AI Evaluation and Testing
  - 5.3.2 Evaluating ML Models
  - 5.3.3 Using AI to Generate Tests
  - 5.3.4 Repeatability
- 5.4 Release
- 5.5 Summary
- 5.6 Discussion Questions
- 5.7 For Further Reading

Chapter 6 System Life Cycle

- 6.1 Design
  - 6.1.1 Co-Design/Development
  - 6.1.2 Using Microservices
  - 6.1.3 Designing for Modifiability
- 6.2 Developing Non-AI Modules
- 6.3 Build
- 6.4 Test
- 6.5 Release and Deploy
  - 6.5.1 Defined versus Continuous Releases
  - 6.5.2 Deployment Patterns
  - 6.5.3 Version Skew

Table of Contents

6.5.4 Matching the Model to Resources

6.6 Operate, Monitor, and Analyze

6.6.1 Monitoring

6.6.2 Incidents

6.6.3 Data Drift

6.6.4 Dynamic Model Updating

6.6.5 Chaos Engineering

6.6.6 Analysis

6.7 Summary

6.8 Discussion Questions

6.9 For Further Reading

Chapter 7 Reliability

7.1 Fundamental Concepts

7.1.1 Under Specified Operating Conditions

7.1.2 Operating Despite Unexpected Inputs

7.1.3 Operating Despite Changes in the Environment

7.1.4 Faults, Failures, and Errors

7.2 Preventing Faults

7.2.1 Model Choice and Preparation

7.2.2 Data Quality, Model Training, and Testing

7.2.3 Feature Engineering

7.2.4 Circuit Breaker

7.2.5 Placement of Responsibilities

7.2.6 Failure Mode and Effects Analysis and Fault Tree Analysis

7.3 Detecting Faults

7.3.1 Total Failure

7.3.2 Questionable Output

7.3.3 Detecting Errors during Operations

7.4 Recovering from Faults

Table of Contents

7.4.1 Redundancy

7.4.2 Managing Error

7.5 Summary

7.6 Discussion Questions

7.7 For Further Reading

Chapter 8 Performance

8.1 Efficiency

8.1.1 Fundamental Concepts of Efficiency

8.1.2 Approaches to Improving Efficiency

8.1.3 Efficiency Considerations of FMs

8.2 Accuracy

8.2.1 Fundamental Concepts of Accuracy

8.2.2 Approaches to Improving Accuracy

8.3 Summary

8.4 Discussion Questions

8.5 For Further Reading

Chapter 9 Security

9.1 Fundamental Concepts

9.1.1 Traditional Definition of Security

9.1.2 Attacks in an AI System

9.2 Approaches to Mitigating Security Concerns

9.2.1 Process Approaches for Enhancing Security

9.2.2 Architectural Approaches for Enhancing Security

9.2.3 Data Preparation Approaches for Enhancing Security

9.2.4 Testing in the Deployment Pipeline

9.2.5 Enhancing Security During System Build

9.2.6 Enhancing Security during Operations

9.2.7 Foundation Models

Table of Contents

9.3 Summary

9.4 Discussion Questions

9.5 For Further Reading

Chapter 10 Privacy and Fairness

10.1 Privacy in AI Systems

10.2 Fairness in AI Systems

10.3 Achieving Privacy

10.3.1 Organizational Approaches

10.3.2 Architecture and Data Preparation

10.3.3 Testing and Operations

10.3.4 Foundation Models

10.4 Achieving Fairness

10.4.1 Organizational Approaches and Tools

10.4.2 Architecture

10.4.3 Data and Model Preparation

10.4.4 Operations

10.4.5 Foundation Models

10.5 Summary

10.6 Discussion Questions

10.7 For Further Reading

Chapter 11 Observability

11.1 Fundamental Concepts

11.2 Evolving from Monitorability to Observability

11.2.1 Monitorability: The Early Stages

11.2.2 Evolution toward Structured Data and Tracing

11.2.3 From Structured Data to Observability

11.3 Approaches for Enhancing Observability

11.3.1 Architecture

Table of Contents

11.3.2 Data Preparation and Model Build
11.3.3 System Build
11.3.4 System Test
11.3.5 Operations
11.4 Summary
11.5 Discussion Questions
11.6 For Further Reading
Chapter 12 The Fraunhofer Case Study: Using a Pretrained Language Model for Tendering
12.1 The Problem Context
12.1.1 Requirements: Interactions, Integration, Data, and Quality Metrics
12.1.2 Development and Operations
12.2 Case Study Description and Setup
12.2.1 MLOps Approach and Generic Architecture
12.2.2 Setup of ML Solution
12.2.3 Stakeholder Involvement
12.2.4 PoC Phase
12.2.5 Implementation Phase
12.2.6 Production Phase
12.3 Summary
12.4 Takeaways
12.5 Discussion Questions
12.6 For Further Reading
Chapter 13 The ARM Hub Case Study: Chatbots for Small and Medium-Size Australian Enterprises
13.1 Introduction
13.2 Our Approach
13.3 LLMs in SME Manufacturing

Table of Contents

13.4 A RAG-Based Chatbot for SME Manufacturing

13.5 Architecture of the ARM Hub Chatbot

13.5.1 ETL/Document Ingestion

13.5.2 Document Preparation and Embedding

13.5.3 Vector Search/Vector Index

13.5.4 Retrieval Using the RAG Chain

13.5.5 User Interface/Response Recording

13.6 MLOps in ARM Hub

13.6.1 What Are the Benefits of MLOps for ARM Hub?

13.6.2 Exploratory Data Analysis

13.6.3 Data Preparation/Ingestion and Prompt Engineering

13.6.4 Model Fine-Tuning

13.6.5 Model Review and Governance

13.6.6 Model Inference and Serving

13.6.7 Model Evaluation

13.7 Ongoing Work

13.8 Summary

13.9 Takeaways

13.10 Discussion Questions

13.11 For Further Reading

Chapter 14 The Banking Case Study: Predicting Customer Churn in Banks

14.1 Customer Churn Prediction

14.1.1 Model Data Sourcing and Validation

14.1.2 Model Feature Engineering and Selection

14.1.3 Model Training and Performance Evaluation

14.1.4 Model Packaging, Deployment, and Inference

14.1.5 Model Performance Monitoring

Table of Contents

14.2 Key Challenges in the Banking Sector

14.2.1 Data Privacy and Security

14.2.2 Traceability and Transparency of AI Decisions

14.3 Summary

14.4 Takeaways

14.5 Discussion Questions

14.6 For Further Reading

Chapter 15 The Future of AI Engineering

15.1 The Shift to DevOps 2.0

15.2 AI's Implications for the Future

15.2.1 The Future of Engineering AI Systems

15.2.2 The Future of MLOps

15.2.3 The Future of Architecture

15.3 AIWare or AI-as-Software

15.4 Trust in AI and the Role of Human Engineers

15.4.1 System-Level Guardrails

15.4.2 Design of Multi-Agent Systems

15.4.3 Understanding and Quality Control of AI Models

15.4.4 Specialization of FMs and Learning Outside the Model

15.5 Summary

15.6 Discussion Questions

15.7 For Further Reading

References

Index

Pearson