Addison-Wesley Data & Analytics Series

SECOND EDITION

Quick Start Guide to Large Language Models

Strategies and Best Practices for ChatGPT, Embeddings, Fine-Tuning, and Multimodal AI

SINAN OZDEMIR

Praise for *Quick Start Guide to Large Language Models*

"By balancing the potential of both open- and closed-source models, *Quick Start Guide to Large Language Models* stands as a comprehensive guide to understanding and using LLMs, bridging the gap between theoretical concepts and practical application."

—Giada Pistilli, Principal Ethicist at Hugging Face

"A refreshing and inspiring resource. Jam-packed with practical guidance and clear explanations that leave you smarter about this incredible new field."

—Pete Huang, author of *The Neuron*

"When it comes to building large language models (LLMs), it can be a daunting task to find comprehensive resources that cover all the essential aspects. However, my search for such a resource recently came to an end when I discovered this book.

"One of the stand-out features of Sinan is his ability to present complex concepts in a straightforward manner. The author has done an outstanding job of breaking down intricate ideas and algorithms, ensuring that readers can grasp them without feeling overwhelmed. Each topic is carefully explained, building upon examples that serve as steppingstones for better understanding. This approach greatly enhances the learning experience, making even the most intricate aspects of LLM development accessible to readers of varying skill levels.

"Another strength of this book is the abundance of code resources. The inclusion of practical examples and code snippets is a game-changer for anyone who wants to experiment and apply the concepts they learn. These code resources provide readers with hands-on experience, allowing them to test and refine their understanding. This is an invaluable asset, as it fosters a deeper comprehension of the material and enables readers to truly engage with the content.

"In conclusion, this book is a rare find for anyone interested in building LLMs. Its exceptional quality of explanation, clear and concise writing style, abundant code resources, and comprehensive coverage of all essential aspects make it an indispensable resource. Whether you are a beginner or an experienced practitioner, this book will undoubtedly elevate your understanding and practical skills in LLM development. I highly recommend *Quick Start Guide to Large Language Models* to anyone looking to embark on the exciting journey of building LLM applications."

—Pedro Marcelino, Machine Learning Engineer,
Co-Founder and CEO @overfit.study

"Ozdemir's book cuts through the noise to help readers understand where the LLM revolution has come from—and where it is going. Ozdemir breaks down complex topics into practical explanations and easy-to-follow code examples."

—Shelia Gulati, Former GM at Microsoft and
current Managing Director of Tola Capital

Quick Start Guide to Large Language Models: Strategies and Best Practices for ChatGPT, Embeddings, Fine-Tuning, and Multimodal AI

Table of Contents

Cover
Title Page
Copyright Page
Contents
Foreword
Preface
Acknowledgments
About the Author
Part I: Introduction to Large Language Models
1 Overview of Large Language Models
What Are Large Language Models?
Definition of LLMs
Popular Modern LLMs
BERT
The GPT Family and ChatGPT
T5
Key Characteristics of LLMs
How LLMs Work
Applications of LLMs
Classical NLP Tasks
Free-Text Generation

Table of Contents

Information Retrieval/Neural Semantic Search

Chatbots

Summary

2 Semantic Search with LLMs

Introduction

The Task

Asymmetric Semantic Search

Solution Overview

The Components

Text Embedder

Document Chunking

Vector Databases

Re-ranking the Retrieved Results

API

Putting It All Together

Performance

The Cost of Closed-Source Components

Summary

3 First Steps with Prompt Engineering

Introduction

Prompt Engineering

Alignment in Language Models

Just Ask

When Just Asking Isnt Enough

Few-Shot Learning

Output Formatting

Prompting Personas

Chain-of-Thought Prompting

Example: Basic Arithmetic

Working with Prompts Across Models

Chat Models versus Completion Models

Coheres Command Series

Table of Contents

- Open-Source Prompt Engineering
- Summary
- 4 The AI Ecosystem: Putting the Pieces Together
  - Introduction
  - The Ever-Shifting Performance of Closed-Source AI
  - AI Reasoning versus Thinking
  - Case Study 1: Retrieval Augmented Generation
    - The Sum of Our Parts: The Retriever and the Generator
    - Evaluating a RAG System
  - Case Study 2: Automated AI Agents
    - Thought Action Observation Response
    - Evaluating an AI Agent
  - Conclusion
- Part II: Getting the Most Out of LLMs
  - 5 Optimizing LLMs with Customized Fine-Tuning
    - Introduction
    - Transfer Learning and Fine-Tuning: A Primer
      - The Fine-Tuning Process Explained
      - Closed-Source Pre-trained Models as a Foundation
    - A Look at the OpenAI Fine-Tuning API
      - The OpenAI Fine-Tuning API
      - Case Study: App Review Sentiment Classification
      - Guidelines and Best Practices for Data
    - Preparing Custom Examples with the OpenAI CLI
    - Setting Up the OpenAI CLI
      - Hyperparameter Selection and Optimization
    - Our First Fine-Tuned LLM
      - Evaluating Fine-Tuned Models with Quantitative Metrics
      - Qualitative Evaluation Techniques
      - Integrating Fine-Tuned OpenAI Models into Applications
      - OpenAI Versus Open-Source Autoencoding BERT

Table of Contents

- Summary

6 Advanced Prompt Engineering

- Introduction
- Prompt Injection Attacks
- Input/Output Validation
  - Example: Using NLI to Build Validation Pipelines
- Batch Prompting
- Prompt Chaining
  - Chaining to Prevent Prompt Stuffing
  - Example: Chaining for Safety Using Multimodal LLMs
- Case Study: How Good at Math Is AI?
  - Our Dataset: MathQA
- Summary

7 Customizing Embeddings and Model Architectures

- Introduction
- Case Study: Building a Recommendation System
  - Setting Up the Problem and the Data
  - Defining the Problem of Recommendation
  - A 10,000-Foot View of Our Recommendation System
  - Generating a Custom Description Field to Compare Items
  - Setting a Baseline with Foundation Embedders
  - Preparing Our Fine-Tuning Data
  - Fine-Tuning Open-Source Embedders Using Sentence Transformers
  - Summary of Results
- Summary

8 AI Alignment: First Principles

- Introduction
- Aligned to Whom and to What End?
  - Instructional Alignment
  - Behavior Alignment
  - Style Alignment

Table of Contents

Value Alignment

Alignment as a Bias Mitigator

The Pillars of Alignment

Data

Training/Tuning Models

Evaluation

Our Three Pillars of Alignment

Constitutional AI: A Step Toward Self-Alignment

Conclusion

Part III: Advanced LLM Usage

9 Moving Beyond Foundation Models

Introduction

Case Study: Visual Q/A

Introduction to Our Models: The Vision Transformer, GPT-2, and DistilBERT

Hidden States Projection and Fusion

Cross-Attention: What Is It, and Why Is It Critical?

Our Custom Multimodal Model

Our Data: Visual QA

The VQA Training Loop

Summary of Results

Case Study: Reinforcement Learning from Feedback

Our Model: FLAN-T5

Our Reward Model: Sentiment and Grammar Correctness

Transformer Reinforcement Learning

The RLF Training Loop

Summary of Results

Summary

10 Advanced Open-Source LLM Fine-Tuning

Introduction

Example: Anime Genre Multilabel Classification with BERT

Using the Jaccard Score to Measure Performance for Multilabel Genre Prediction of

Table of Contents

Anime Titles

A Simple Fine-Tuning Loop

General Tips for Fine-Tuning Open-Source LLMs

Summary of Results

Example: LaTeX Generation with GPT2

Prompt Engineering for Open-Source Models

Summary of Results

Sinans Attempt at Wise Yet Engaging Responses: SAWYER

Step 1: Supervised Instruction Fine-Tuning

Step 2: Reward Model Training

Step 3: Reinforcement Learning from (Estimated) Human Feedback

Summary of Results

Updating Our LLM with Fresh Knowledge

Summary

11 Moving LLMs into Production

Introduction

Deploying Closed-Source LLMs to Production

Cost Projections

API Key Management

Deploying Open-Source LLMs to Production

Preparing a Model for Inference

Interoperability

Quantization

Knowledge Distillation

Cost Projections with LLMs

Pushing to Hugging Face

Summary

12 Evaluating LLMs

Introduction

Evaluating Generative Tasks

Generative Multiple Choice

Free Text Response

Table of Contents

Benchmarking

Evaluating Understanding Tasks

Embeddings

Calibrated Classification

Probing LLMs for a World Model

Conclusion

Keep Going!

Part IV: Appendices

A LLM FAQs

B LLM Glossary

C LLM Application Archetypes

Index