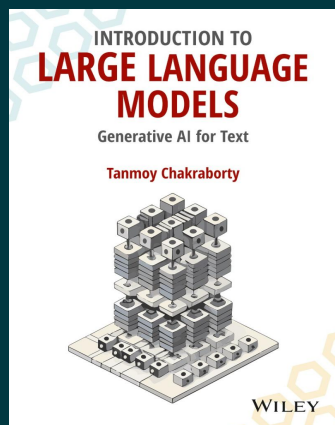




Introduction to Large Language Models

By [Tanmoy Chakraborty](#)

Paperback

9789363864740

₹899.00

Pages: 484

Description

Introduction to Large Language Models (LLMs) is a comprehensive guide for understanding the foundations and advancements of Generative AI for Text. Designed for educators and enthusiasts, the book starts with key linguistic concepts and progresses through NLP fundamentals—from word embeddings to pretrained foundational models.

Readers will learn how LLMs process and generate language, overcome limitations, and enhance performance using techniques like prompt engineering, retrieval-augmented generation, and human alignment. The book uniquely presents cutting-edge research in a concise format, enriched with visual aids, exercises, and practical resources.

Ideal for computer science faculty, this resource offers both theoretical insights and real-world applications, showcasing how LLMs like ChatGPT are transforming technology and advancing AI innovation.

About the Author

Tanmoy Chakraborty

Dr. Tanmoy Chakraborty is an Associate Professor in the Department of Electrical Engineering at IIT Delhi and an Associate Faculty Member at the Yardi School of Artificial Intelligence. An ACM Distinguished Speaker (2023–2025) and former Ramanujan Fellow (2018–2023), he has held key academic roles, including heading the Infosys Centre for Artificial Intelligence at IIIT Delhi.

Dr. Chakraborty earned his Ph.D. as a Google India scholar at IIT Kharagpur and completed a postdoctoral fellowship at the University of Maryland, College Park. His research spans Natural Language Processing (NLP), Graph Neural Networks, and Social Computing, with a focus on creating frugal, explainable LLMs for applications in mental health and cyber-informatics.

He leads the Laboratory for Computational Social Systems (LCS2) and also recipient of multiple faculty awards from Google, Adobe, and Accenture,

Table of Contents

Endorsement

Preface

Acknowledgement

Foreword

1 Introduction

1.1 What is a Language Model?

1.2 Evolution of Language Modelling Technologies

1.3 Scaling Laws in Language Models

1.4 Evolution of LLMs

1.4.1 The Emergence and Development of LLMs

1.4.2 Implications of Encoder-Decoder in LLM Development

1.4.3 Optimising Scale and Resource Efficiency in LLMs

1.5 Organisation of the Book

Additional Resources

Bibliography

2 An Overview of Natural Language Processing and Neural Networks

Part I: Natural Language Processing

2.1 Computational Linguistics and Natural Language Processing

2.2 Overview of the Natural Language Processing Pipeline

2.3 Morphology

2.3.1 Morphemes

2.3.2 Stemming

2.3.3 Lemmatisation

2.3.4 Lexicon

2.4 Tokenisation

2.4.1 Advanced Techniques: Subword Tokenisation

2.5 Syntactics

2.6 Semantics

2.7 Introduction to Language Modelling

Part II: Neural Networks

2.8 The Perceptron

2.8.1 Definition

2.8.2 Implementing AND, OR, and XOR Logic

2.9 Multilayer Perceptron

2.9.1 Neural Networks

2.9.2 Types of Activation Functions

2.10 Training Neural Networks

2.10.1 Backpropagation

2.10.2 Batching

2.10.3 Hyperparameters

2.10.4 Regularisation

2.11 Vanishing and Exploding Gradients

2.12 Evaluation Metrics

2.13 Summary

Additional Resources

Exercises

Bibliography

3 Word Embedding

3.1 Distributional Hypothesis

3.2 Vector Semantics

3.2.1 Defining and Measuring Semantic Similarity

3.3 Types of Word Embedding

3.3.1 Frequency-Based Embeddings

3.3.2 Word2Vec

3.3.3 Global Vectors for Word Representation

3.3.4 FastText

3.4 Bias in Word Embedding

3.5 Limitations of Word Embedding Methods

3.6 Applications of Word Embeddings

3.7 Summary

Additional Resources

Exercises

Bibliography

4 Statistical Language Model

4.1 Statistical Language Model

4.1.1 The Conditional Probability

4.1.2 The Chain Rule of Probability

4.1.3 The Markov Assumption

4.1.4 Unigram Language Model

4.1.5 Bigram Language Model

4.2 Smoothing

4.2.1 The Unknown Tokens

4.2.2 Smoothing

4.2.3 Back-Off

4.2.4 Interpolation

4.2.5 Good-Turing

4.3 Evaluation of Language Model

4.3.1 Extrinsic Evaluation

4.3.2 Intrinsic Evaluation

4.3.3 Human Evaluation

4.3.4 Evaluation Metrics

4.3.5 Benchmark Suits

4.4 Limitations of Statistical Language Models

4.5 Summary

Additional Resources

Exercises

Bibliography

5 Neural Language Models

5.1 Convolutional Neural Networks

5.1.1 Components of CNNs: Kernel, Stride, Pooling, and Padding

5.1.2 Hierarchical and Dilated Convolutions

5.1.3 Applications of CNNs in NLP

5.2 Recurrent Neural Networks

5.2.1 Training RNNs

5.2.2 Applications of RNNs

5.2.3 Challenges in Sequence Modelling

5.2.4 RNN Variants: LSTM, GRU, and Bidirectional RNNs

5.3 Sequence-to-Sequence Models

5.3.1 Training Sequence-to-Sequence Models

5.3.2 Inference Decoding

5.3.3 Applications of Sequence-to-Sequence Models

5.4 Attention Mechanisms

5.4.1 Introduction to Attention

5.4.2 Advantages of Attention

5.4.3 Variants of Attention

5.5 Limitations of Neural Language Models

5.6 Summary

Additional Resources

Exercises

Bibliography

6 Transformers

6.1 Self-Attention

6.1.1 Multi-Head Self-Attention

6.2 Transformer Encoder Block

6.2.1 Components of the Transformer Encoder Block

6.2.2 Feed-Forward Neural Network

6.2.3 Layer Normalisation

6.2.4 Residual Connections

6.3 Transformer Decoder Block

6.3.1 Masked Multi-Head Self-Attention

6.3.2 Cross-Attention (Encoder-Decoder Attention)

6.4 Positional Embeddings

6.4.1 Types of Positional Embeddings

6.4.2 Rotary Position Embedding

6.5 Efficient Attention Mechanisms

6.5.1 KV Caching in Multi-Head Self-Attention

6.5.2 Multi-Query Attention

6.5.3 Grouped-Query Attention

6.5.4 Sliding Window Attention

6.6 An Alternate Formulation of Transformers

6.6.1 Residual Stream Perspective of Transformers

6.6.2 Attention Heads: Reading and Writing

6.6.3 Feed-Forward Networks: Transformation of Residual Streams

6.6.4 Prediction Head: Generating the Next Token

6.6.5 Decomposing the Transformer: Attention and Feed-Forward Contributions

6.6.6 Residual Networks as Shallow Ensembles

6.6.7 Interpreting the Mechanism of LLMs

6.7 Summary

Additional Resources

Exercises

Bibliography

7 Language Model Pretraining

7.1 Embeddings from Language Model

7.1.1 Architecture and Training of ELMo

7.1.2 Applications of ELMo

7.1.3 Limitations of ELMo

7.2 Evaluation Datasets

7.3 Encoder-Based Pretraining

7.3.1 Fundamentals of Encoder-Based Models

7.3.2 Training Paradigm

7.3.3 BERT Pretraining

7.3.4 Applications and Limitations

7.4 Decoder-Based Pretraining

7.4.1 Decoder-Based Architecture

7.4.2 Training Paradigm

7.4.3 GPT Pretraining

7.4.4 Applications and Limitations

7.5 Encoder-Decoder Based Pretraining

7.5.1 Architecture

7.5.2 Joint Pretraining Strategy

7.5.3 T5 Pretraining

7.5.4 Applications and Limitations

7.6 Emergence of Large Language Models

7.7 Limitations of Pretraining

7.8 Summary

Additional Resources

Exercises

Bibliography

8 Fine-Tuning and Alignment of LLMs

8.1 Moving from Pretraining to Fine-Tuning

8.2 Fine-Tuning on Various Task-Specific Applications

8.2.1 Sequence Classification

8.2.2 Pairwise Sequence Classification

8.2.3 Sequence Labelling

8.2.4 Learning Spans

8.2.5 Challenges in Classical Fine-Tuning Methods

8.3 Instruction Tuning

8.4 Alignment Methods

8.4.1 Reinforcement Learning from Human Feedback

8.4.2 Direct Preference Optimisation

8.5 Summary

Additional Resources

Exercises

Bibliography

9 Prompting Strategies in LLMs

9.1 Prompt Engineering

9.1.1 Prompt Shape

9.1.2 Manual Template Engineering

9.1.3 Automated Template Learning

9.1.4 Continuous Prompts

9.2 Prompt Application

9.2.1 In-Context Learning

9.2.2 Knowledge Probing

9.2.3 Classification-Based Tasks

9.2.4 Information Extraction

9.2.5 Reasoning in Natural Language Processing

9.2.6 Question Answering

9.2.7 Text Generation

9.2.8 Automatic Evaluation of Text Generation

9.3 Chain-of-Thoughts

9.4 Tree-of-Thoughts

9.5 Graph-of-Thoughts

9.6 Summary

Additional Resources

Exercises

Bibliography

10 Efficient Methods for Fine-Tuning LLMs

10.1 Model Compression with Knowledge Distillation

10.1.1 White-Box Knowledge Distillation

10.1.2 Meta Knowledge Distillation

10.1.3 Black-Box Knowledge Distillation

10.2 Model Compression Techniques

10.2.1 Model Pruning

10.2.2 Model Quantisation

10.3 Parameter-Efficient Fine-Tuning

10.3.1 Adapters

10.3.2 Prefix Tuning

10.3.3 Prompt Tuning

10.3.4 Selective PEFT Techniques

10.3.5 Reparameterisation-Based PEFT Techniques

10.3.6 Hybrid Approaches for Efficient Fine-Tuning

10.4 Efficient Strategies for Fine-Tuning LLMs

10.4.1 Mixed-Precision Tuning

10.4.2 Data Selection for Efficient Fine-Tuning

10.4.3 Prompt Compression

10.5 Summary

Additional Resources

Exercises

Bibliography

11 Augmented Large Language Models

11.1 Retrieval-Augmented Generation

11.1.1 Indexing in RAGs

11.1.2 Context Searching in RAGs

11.1.3 Prompting in RAGs

11.1.4 Inferencing in RAGs

11.1.5 Comparison of RAGs with LLMs

11.2 Evaluation of RAGs

11.2.1 Assessing of Retrieval Quality

11.2.2 Generation Quality

11.2.3 Knowledge Integration and Factuality Evaluation

11.2.4 Response Time and Efficiency

11.2.5 User Satisfaction

11.2.6 RAGAs Framework for RAG Evaluation

11.3 Tool Calling with LLMs

11.3.1 Autonomously Determining Which Tools to Use and Where

11.3.2 Examples of Different Tools

11.3.3 Evaluation of Code Generation Capabilities of Agents

11.3.4 Error Handling and Optimisation

11.4 LLM Augmentation with Agents

11.4.1 Reasoning in LLM Agents

11.4.2 Planning in LLM Agents

11.4.3 Handling Memory in LLM Agents

11.5 Summary

Additional Resources

Exercises

Bibliography

12 Multilingual and Multimodal LLMs

12.1 Multilingual Language Models

12.1.1 The Evolution of Multilingual NLP

12.1.2 The Need for Multilingual LLMs

12.1.3 Cross-Lingual Representation Learning

12.1.4 Applications

12.2 Multimodal Language Models

12.2.1 Integration of Diverse Modalities

12.2.2 Applications

12.3 Training Multilingual and Multimodal LLMs

12.3.1 Efficient Data Collection and Preprocessing

12.3.2 Model Training Strategies

12.4 Addressing Challenges in Multilingual and Multimodal LLMs

12.4.1 Challenges in Multilingual LLMs

12.4.2 Challenges in Multimodal LLMs

12.5 Future Directions and Emerging Trends

12.6 Limitations of Multilingual and Multimodal LLMs

12.7 Summary

Additional Resources

Exercises

Bibliography

13 Responsible LLMs

13.1 Inaccurate, Inappropriate, and Unethical Behaviour of LLMs

13.2 Responsible AI

13.3 Bias

13.3.1 Visibility of Bias

13.3.2 Source of Bias

13.4 Bias Mitigation

13.5 Summary

Additional Resources

Exercises

Bibliography

14 Advanced Topics in Large Language Models

14.1 Reasoning with LLMs

14.1.1 Advancements in Reasoning Capabilities

14.1.2 Challenges in Reasoning with LLMs

14.1.3 Types of Reasoning Tasks

14.1.4 How Do LLMs Approach Reasoning?

14.1.5 Evaluating Reasoning Abilities in LLMs

14.2 Handling Long Context in LLMs

14.2.1 Challenges in Processing Long Context

14.2.2 Training and Fine-Tuning Approaches to Extend Context Length

14.2.3 Evaluation of Long-Context LLMs

14.3 Model Editing

14.3.1 Conditions for Successful Editing

14.3.2 Methods for Model Editing

14.3.3 Metrics for Evaluation of Model Editing

14.4 Hallucination in LLMs

14.4.1 Definition

14.4.2 Sources of Hallucination

14.4.3 Metrics Measuring Hallucination

14.4.4 Hallucination Mitigation

14.5 Self-Evolving LLMs

14.5.1 Conceptual Framework

14.5.2 Evolution Objectives and Techniques

14.5.3 Challenges

14.6 Summary

Additional Resources

Exercises

Bibliography

15 LLMs in Action

15.1 An Overview of the Landscape

15.1.1 Tracing the Evolution and Importance of LLMs in Contemporary AI

15.1.2 Open-Source vs Closed-Source Paradigms: Benefits and Trade-offs

15.2 A Panoramic View of LLMs

15.2.1 General-Purpose Large Language Models

15.2.2 Language-Specific LLMs

15.2.3 Domain-Specific LLMs

15.2.4 Task-Specific LLMs

15.3 Diverse Applications of LLMs

15.3.1 Healthcare: Enhancing Diagnostics and Patient Care

15.3.2 Finance: Transforming Data Analysis and Risk Management

15.3.3 Legal: Streamlining Research and Case Management

15.3.4 Education: Personalised Learning and Academic Support

15.4 Emerging Trends and Future Directions in LLMs

15.4.1 Beyond Text: The Advent of Multimodal LLMs

15.4.2 Autonomous Agents: The LLM Leap in AI Evolution (AutoGPT)

15.5 Summary

Additional Resources

Exercises

Bibliography

Index

To purchase this product, please visit:

<https://www.wileyindia.com/introduction-to-large-language-models.html>



Scan to buy