

brief contents

- 1 ■ Understanding large language models 1
- 2 ■ Working with text data 17
- 3 ■ Coding attention mechanisms 50
- 4 ■ Implementing a GPT model from scratch to generate text 92
- 5 ■ Pretraining on unlabeled data 128
- 6 ■ Fine-tuning for classification 169
- 7 ■ Fine-tuning to follow instructions 204
- A ■ Introduction to PyTorch 251
- B ■ References and further reading 289
- C ■ Exercise solutions 300
- D ■ Adding bells and whistles to the training loop 313
- E ■ Parameter-efficient fine-tuning with LoRA 322

contents

preface xi
acknowledgments xiii
about this book xv
about the author xix
about the cover illustration xx

1 *Understanding large language models* 1

- 1.1 What is an LLM? 2
- 1.2 Applications of LLMs 4
- 1.3 Stages of building and using LLMs 5
- 1.4 Introducing the transformer architecture 7
- 1.5 Utilizing large datasets 10
- 1.6 A closer look at the GPT architecture 12
- 1.7 Building a large language model 14

2 *Working with text data* 17

- 2.1 Understanding word embeddings 18
- 2.2 Tokenizing text 21
- 2.3 Converting tokens into token IDs 24
- 2.4 Adding special context tokens 29