

Contents

Part I Fundamentals of Foundation Models

1	Foundation Models and Generative AI	3
1.1	What Are Foundation Models and Generative AI?	3
1.2	Foundations Models at Scale	5
1.2.1	Neural Scaling Laws	5
1.2.2	Emerging Abilities	6
1.3	Lifecycle of Foundation Models	7
1.3.1	Data Preparation	8
1.3.2	Model Training	8
1.3.3	Model Deployment	9
1.4	Canonical Examples of Foundation Models	10
1.5	Our Perspective on Foundation Models	11
2	Neural Networks	13
2.1	Introduction	13
2.2	Basic Neural Network Modules	14
2.3	Transformers	16
2.3.1	Token Embedding and Position Encoding	16
2.3.2	Attention	17
2.3.3	Transformer Types	19
2.4	Large Language Models	20
2.4.1	Next-Token Prediction	20
2.4.2	Decoding Strategies	21
2.4.3	Alignment Strategies	22
2.4.4	Parameter-Efficient Fine-Tuning	23
3	Learning and Generalization of Vision Transformers	25
3.1	Introduction	25
3.2	Background and Related Work	26
3.2.1	Problem Formulation and Learning Algorithm	27
3.3	Theoretical Characterization of Transformers	29

3.3.1	Main Theoretical Insights	29
3.3.2	Data Model.....	30
3.3.3	Formal Theoretical Results.....	31
3.4	Performance Evaluation.....	33
4	Formalizing In-Context Learning in Transformers.....	37
4.1	Introduction.....	37
4.2	Background and Related Work	39
4.2.1	Formalizing In-Context Learning with Transformers.....	40
4.3	Theoretical Characterization of In-Context Learning	43
4.3.1	Main Theoretical Insights	43
4.3.2	The Modeling of Training Data and Tasks	43
4.3.3	In-Domain and Out-of-Domain Generalization with Sample Complexity Analysis.....	45
4.3.4	ICL With Magnitude-Based Model Pruning	47
4.3.5	The Mechanism of ICL by the Trained Transformer	48
4.4	Performance Evaluation.....	50
 Part II Advanced Topics in Foundation Models		
5	Automated Visual Prompting	55
5.1	Introduction.....	55
5.2	Background and Related Work	57
5.3	AutoVP Framework	59
5.3.1	Input Scaling	60
5.3.2	Visual Prompt.....	60
5.3.3	Pre-trained Classifier	61
5.3.4	Output Label Mapping	61
5.3.5	End-to-End Hyper-Parameter Tuning	63
5.4	Performance Evaluation.....	64
6	Prompting Large Language Models with Privacy	67
6.1	Introduction.....	67
6.2	Background and Related Work	68
6.2.1	Differential Privacy (DP).....	68
6.2.2	Document Processing with Privacy	70
6.3	DP-Prompt.....	71
6.4	Performance Evaluation.....	73
7	Memory-Efficient Fine-Tuning for Foundation Models	77
7.1	Introduction.....	77
7.2	Algorithmic Foundations of ZO Optimization	78
7.3	Applying ZO Optimization for Memory-Efficient Fine-Tuning ...	81
8	Large Language Models Meet Time Series	87
8.1	Introduction.....	87
8.2	Background and Related Work	89

8.3	Time-LLM	90
8.3.1	Model Structure	91
8.4	Performance Evaluation	94
9	Large Language Models Meet Speech Recognition	99
9.1	Introduction	99
9.2	Background and Related Work	101
9.2.1	ASR Rescoring and Error Correction	101
9.2.2	Noise-Robust ASR	102
9.2.3	HyPoradise (HP) Benchmarks	102
9.3	Noise-Aware Generative Error Correction	103
9.3.1	Language-Space Noise Embedding	104
9.3.2	Audio Noise Distillation	105
9.4	Performance Evaluation	106
10	Benchmarking Foundation Models Using Synthetic Datasets	109
10.1	Introduction	109
10.2	Background and Related Work	112
10.3	SynBench	113
10.3.1	Synthetic Data	113
10.3.2	Main Theorem	114
10.3.3	Objective	116
10.3.4	Robustness-Accuracy Quantification	117
10.4	Performance Evaluation	118
11	Machine Unlearning for Foundation Models	123
11.1	Introduction	123
11.2	Research Objective, Formulation, and Related Work	124
11.3	Sparse Optimization for MU: Leveraging Model Sparsity for Efficient and Effective Unlearning	129
11.4	Second-Order Optimization for MU: Iterative Influence-Guided Unlearning	134
11.5	Adversarial Evaluation of MU	137
Part III Trust and Safety in Foundation Models		
12	Trustworthiness Evaluation of Large Language Models	149
12.1	Introduction	149
12.2	Background and Related Work	152
12.2.1	Large Language Models (LLMs)	152
12.2.2	Evaluation on LLMs	154
12.2.3	Trustworthiness-Related Benchmarks	155
12.3	Guidelines and Principles for Trustworthiness Assessment of LLMs	155
12.3.1	Truthfulness	156
12.3.2	Safety	157
12.3.3	Fairness	158

12.3.4	Robustnesss	158
12.3.5	Privacy	159
12.3.6	Machine Ethics	160
12.3.7	Transparency	160
12.3.8	Accountability	161
12.4	Main Insights from TrustLLM Evaluation	162
12.4.1	Overall Observations	162
12.4.2	Novel Insights into Individual Dimensions of Trustworthiness	164
13	Attacks and Defenses on Aligned Large Language Models	167
13.1	Introduction	167
13.2	Background and Related Work	169
13.3	Gradient Cuff	170
13.3.1	Refusal Loss Function and Landscape Exploration	170
13.3.2	Gradient Norm Estimation	172
13.3.3	Gradient Cuff: Two-Step Jailbreak Detection	173
13.3.4	Performance Evaluation	174
13.4	Defensive Prompt Patch	176
13.4.1	Preliminaries	176
13.4.2	Score Evaluation	177
13.4.3	DPP Training Algorithm	178
13.4.4	Performance Evaluation	182
14	Safety Risks in Fine-Tuning Large Language Models	185
14.1	Introduction	185
14.2	Background and Related Work	187
14.3	Performance Evaluation	189
14.3.1	Experiment Setup	189
14.3.2	Numerical Results	190
15	Watermarks for Large Language Models	195
15.1	Introduction	195
15.2	Background and Related Work	196
15.3	Duwak: Dual Watermarking for LLMs	198
15.3.1	Token Probability Watermark	199
15.3.2	Contrastive Search Watermark	199
15.3.3	Watermark Detection in Duwak	200
15.4	Performance Evaluation	202
15.4.1	Experiment Setup	202
15.4.2	Numerical Results	203
16	AI-Generated Text Detection	209
16.1	Introduction	209
16.2	Background and Related Work	210
16.3	RADAR: Robust AI-Text Detection Using Adversarial Learning	212

16.3.1	Training Paraphraser via Clipped PPO with Entropy Penalty	213
16.3.2	Training Detector via Reweighted Logistic Loss	214
16.3.3	RADAR Algorithm	215
16.4	Performance Evaluation.....	215
16.4.1	Experiment Setup.....	215
16.4.2	Performance Evaluation and Comparison with Existing Methods	218
16.4.3	AI-Text Detection Transferability of RADAR	219
17	Backdoor Risks in Diffusion Models	221
17.1	Introduction.....	221
17.2	Background and Related Work	223
17.3	VillanDiffusion: A Unified Backdoor Attack Framework.....	224
17.3.1	Backdoor Unconditional Diffusion Models as a Distribution Mapping Problem.....	224
17.3.2	Generalization to Various Schedulers.....	225
17.3.3	Generalization to ODE and SDE Samplers.....	227
17.3.4	Unified Loss Function for Unconditional Generation with Image Triggers	228
17.3.5	Generalization to Conditional Generation	229
17.4	Backdoor Detection and Mitigation for Diffusion Models	230
17.5	Performance Evaluation.....	231
17.5.1	Experiment Setup.....	231
17.5.2	Caption-Trigger Backdoor Attacks on Text-to-Image DMs.....	232
17.5.3	Image-Trigger Backdoor Attacks on Unconditional DMs.....	234
18	Prompt Engineering for Safety Red-Teaming: A Case Study on Text-to-Image Diffusion Models	237
18.1	Introduction.....	237
18.2	Background and Related Work	238
18.3	Prompting4Debugging (P4D).....	241
18.4	Ring-A-Bell	243
18.5	Performance Evaluation.....	245
18.5.1	P4D Results	246
18.5.2	Ring-A-Bell Results	249
	References.....	255
	Index.....	309