

Content	1 Introduction	1.1 The Shift from Cloud to Edge
		1.2 Drivers Behind Edge LLM
	2 Preliminaries and Essentials	2.1 Evolution of Basic Model Architecture
		2.2 Performance Metrics and Bottlenecks
		2.3 Characteristics of Edge Devices
	3 Algorithm	3.1 Model Compression Techniques
		3.2 Meta-Architecture Design
	4 Framework	4.1 End-to-End Frameworks and Backends
		4.2 High-Speed Computation Kernels
		4.3 Graph Optimization
		4.4 Memory Optimization
		4.5 Pipeline Optimization
		4.6 Multi-device Collaboration
	5 Hardware	5.1 ASIC & FPGA
	5.2 PIM	
6 Applications and Benchmark	6.1 Advances of Hardware Development	
	6.2 Applications and Task Scenarios	
	6.3 Native Deployment Framework	
7 Future Trends	7.1 Research Frontiers	
	7.2 Community Ecosystem	