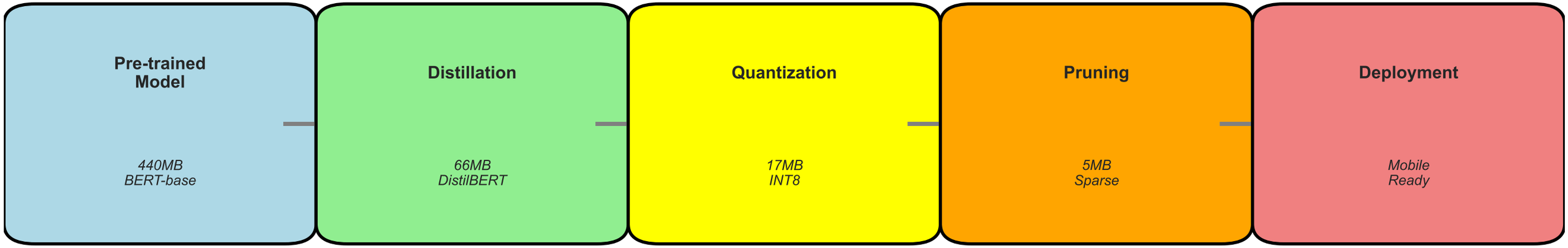
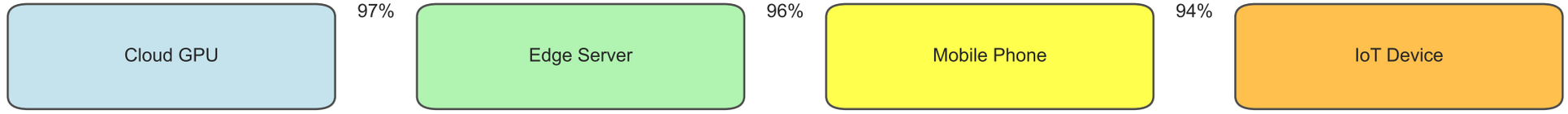
Model Deployment Pipeline: From Research to Production



Size:	440MB	66MB	17MB	5MB	5MB
Latency:	100ms	80ms	40ms	20ms	15ms
Accuracy:	100%	97%	96%	94%	94%

Deployment Targets:



Each stage trades model size for deployment flexibility