Type	Optimization	Threat	FT	Orca	vLLM	FlexGen	FastServe	FastServe2	Sarathi	Lookahead	REST	SpecInfer	Medusa	DistServe	Splitwise	LoongServe	TetriInfer	InfiniteLLM
			22	22	23	23	23	23	23	24	24	24	24	24	24	24	24	24
Batch	Iteration-Level Batch			Initial	✓		✓		✓					✓			✓	
	Chunked Prefill								Initial								✓	
	Prepack Prefill											✓		✓			✓	
Parallel	Speculation	**S**								✓	✓	✓	✓					
	Prompt-Based Speculation	**S**								✓								
	Context-Based Speculation	**S**									✓							
	Tensor Parallelism		✓															
	SafeTensors																	
	Sequence Parallelism															✓		
Memory	Paging				Initial				✓								✓	
	Multi-Query Attention																	
	Grouped-Query Attention	**T**																
Tranmission	Offloading	**SE**				✓	✓											
	Duplication	**T**																
	Pulling	**SET**												✓				
	Request Migration															✓		
	Disaggregated Arch													✓	✓		✓	
Scheduling	Priority-Based	**T**					✓		✓								✓	
	Request-Level Prediction	**T**					✓					Small Model					✓	
	Machine-level Scheduler	**ET**					✓								✓	✓	✓	✓
	Instance Flip														✓		✓	
	Global Profiling	**P**				✓								✓	✓			
Verification	Open Source	**V**															✓	