GPU Memory Usage Summary Across RAG Systems

System	LLM Model (GB)	Encoder (GB)	Vector DB (GB)	KV Cache (GB)	Total Memory (GB)	Fragmentation (%)	Swapping Overhead (ms)	Memory Efficiency (%)
Classical RAG	16.0	0.44	1.5	4.0	21.9	42.9	350	70%
PipeRAG	16.0	0.44	1.5	4.0	21.9	33.3	200	75%
FlashRAG	16.0	0.44	1.5	3.0	20.9	25.0	150	80%
EdgeRAG	8.0	0.22	0.5	1.5	10.2	17.6	100	85%
CSD-Enhanced RAG	16.0	0.00	0.0	4.0	20.0	5.3	50	95%