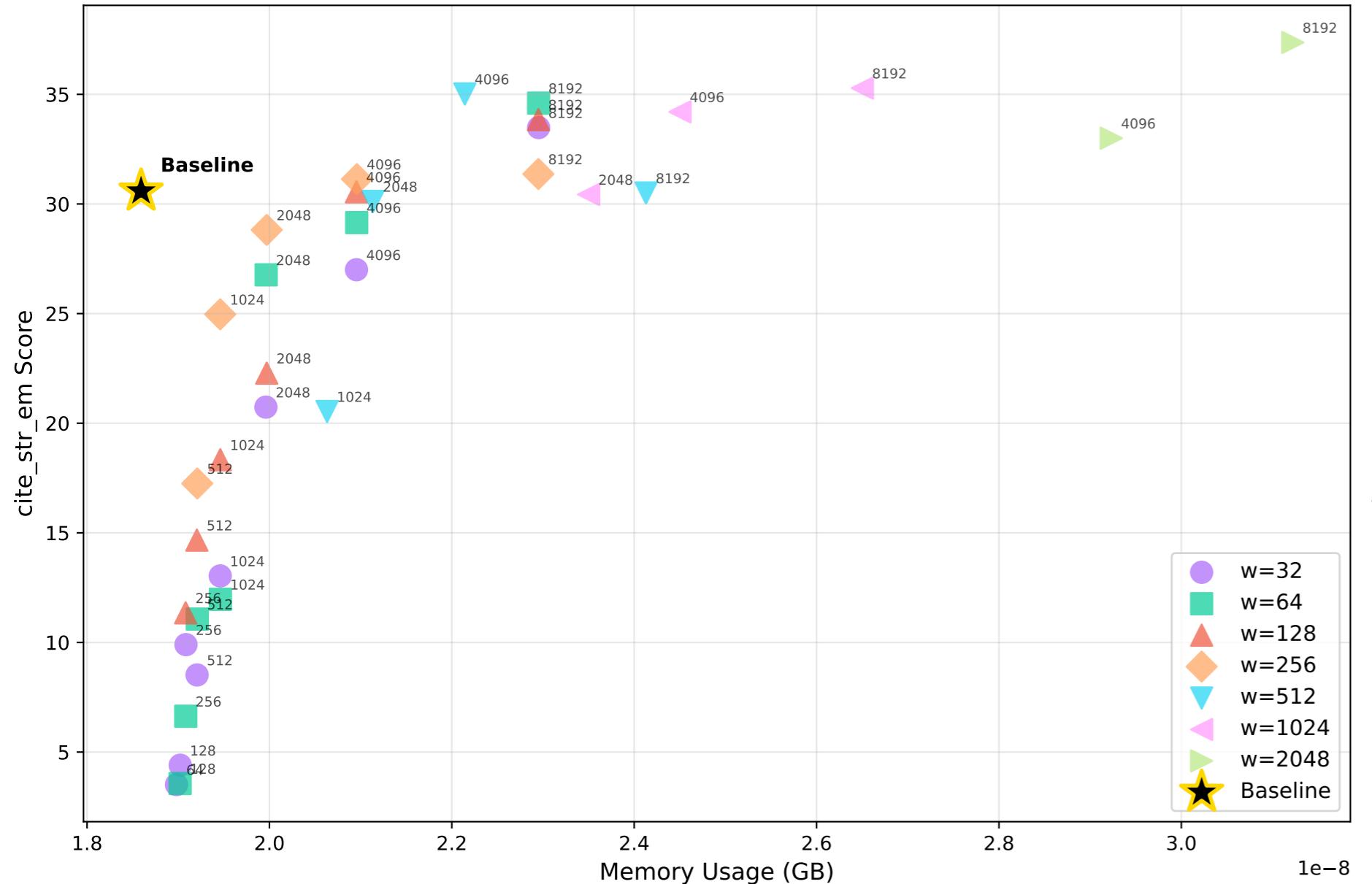


PyramidKV (k=5): cite_str_em vs Memory
(DeepSeek-R1-Distill-Llama-8B, 16k)



PyramidKV (k=5): cite_str_em vs Latency
(DeepSeek-R1-Distill-Llama-8B, 16k)

