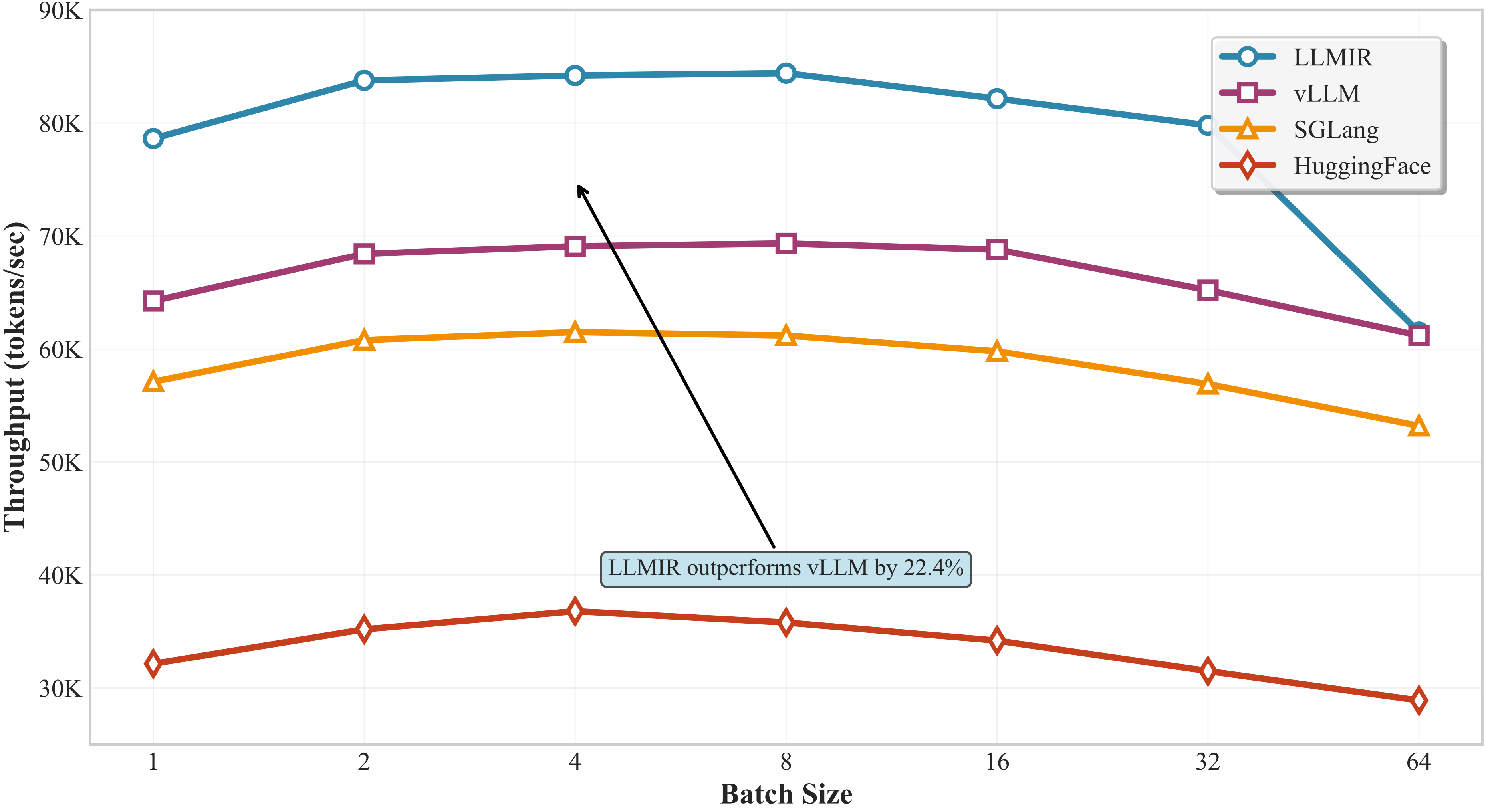
LLM Inference Performance Comparison (LLaMA-2-13B)


Throughput (tokens/sec)
90K
80K
70K
60K
50K
40K
30K
Batch Size
LLMIR
vLLM
SGLang
HuggingFace
LLMIR outperforms vLLM by 22.4%