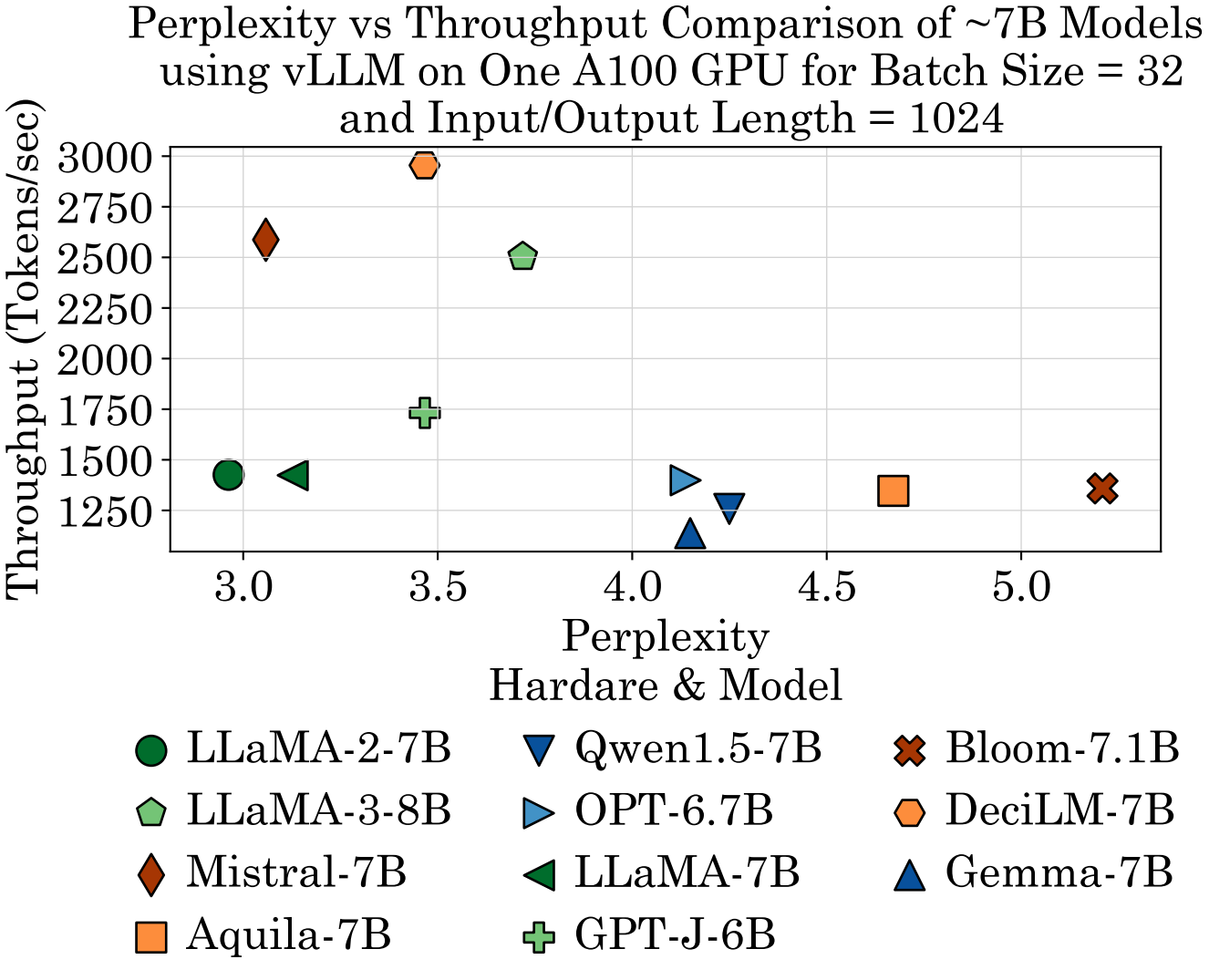
Perplexity vs Throughput Comparison of ~7B Models
using vLLM on One A100 GPU for Batch Size = 32
and Input/Output Length = 1024
Throughput (Tokens/sec)
3000
2750
2500
2250
2000
1750
1500
1250
3.0
3.5
4.0
4.5
5.0
Perplexity
Hardare & Model
LLaMA-2-7B
Qwen1.5-7B
Bloom-7.1B
LLaMA-3-8B
OPT-6.7B
DeciLM-7B
Mistral-7B
LLaMA-7B
Gemma-7B
Aquila-7B
GPT-J-6B