

■ H100 TRT-LLM LLaMA-2-7B ♦ A100 vLLM LLaMA-2-7B

- ▶ H100 TRT-LLM LLaMA-3-8B
 ▶ A100 TRT-LLM LLaMA-2-7B
 ▶ H100 vLLM LLaMA-3-8B
 ▶ A100 vLLM LLaMA-3-8B
- ♦ A100 DS-MII LLaMA-2-7B ♦ A100 TRT-LLM LLaMA-3-8B