

# LLM prefill latency (ms)

Model Size

1B

7.00

2.51

0.80

0.25

3B

16.89

6.06

1.92

0.56

13B

60.46

21.70

6.90

2.00

V100

A100

H100

B100

Inference GPU

50

25