

LLaMA		FP16	H800	PPL.NN.LLM		Rev0+
model size(B)	gpu	batch	input len	output len	prefill(ms)	
7	1	1	1	8	256	6.78
7	1	2	2	8	256	6.84
7	1	4	4	8	256	7.12
7	1	8	8	8	256	7.10
7	1	16	16	8	256	7.94
7	1	32	32	8	256	10.41
7	1	64	64	8	256	16.36
7	1	128	128	8	256	30.03
7	1	256	256	8	256	58.69
7	1	384	384	8	256	83.42
7	1	512	512	8	256	110.82
7	1	768	768	8	256	165.19
7	1	1	1024	1024	1024	30.86
7	1	2	1024	1024	1024	58.33
7	1	4	1024	1024	1024	109.13
7	1	8	1024	1024	1024	213.48
7	1	16	1024	1024	1024	415.95
7	1	32	1024	1024	1024	830.35
7	1	64	1024	1024	1024	1633.89
7	1	80	1024	1024	1024	2078.37
13	2	1	1	8	256	8.24
13	2	2	2	8	256	8.43
13	2	4	4	8	256	8.53
13	2	8	8	8	256	9.78
13	2	16	16	8	256	11.10
13	2	32	32	8	256	13.74
13	2	64	64	8	256	21.33
13	2	128	128	8	256	40.17
13	2	256	256	8	256	69.08
13	2	384	384	8	256	99.10
13	2	512	512	8	256	131.19
13	2	1024	1024	8	256	246.24
13	2	1	1024	1024	1024	41.19
13	2	2	1024	1024	1024	70.31
13	2	4	1024	1024	1024	127.86
13	2	8	1024	1024	1024	243.83
13	2	16	1024	1024	1024	470.57
13	2	32	1024	1024	1024	945.02
13	2	64	1024	1024	1024	1901.01
13	2	96	1024	1024	1024	2875.81
65	8	1	1	8	256	16.38
65	8	2	2	8	256	17.00
65	8	4	4	8	256	16.43
65	8	8	8	8	256	18.80
65	8	16	16	8	256	20.83
65	8	32	32	8	256	29.38
65	8	64	64	8	256	44.65
65	8	128	128	8	256	75.48
65	8	256	256	8	256	136.26
65	8	384	384	8	256	200.19

65	8	512	8	256	264.19
65	8	1024	8	256	496.59
65	8	1	1024	1024	76.17
65	8	2	1024	1024	136.40
65	8	4	1024	1024	265.35
65	8	8	1024	1024	494.73
65	8	16	1024	1024	961.91
65	8	32	1024	1024	1903.35
65	8	64	1024	1024	3819.21
65	8	96	1024	1024	5808.93
65	8	128	1024	1024	7733.57

+GEMM 32MiB Workspae

decode min(ms)	decode max(ms)	decode avg(ms)	total avg(ms)	throughput(tokens/s)
6.65	7.01	6.83	6.86	145.85
6.71	7.11	6.91	6.94	288.32
6.73	7.16	6.95	6.97	573.66
6.81	7.28	7.05	7.07	1131.10
6.88	7.41	7.15	7.18	2229.65
7.39	8.56	7.98	8.02	3992.18
7.58	9.52	8.55	8.61	7429.85
8.68	12.25	10.47	10.58	12095.66
11.67	18.72	15.20	15.42	16597.23
17.71	28.08	22.90	23.22	16536.86
18.77	32.12	25.45	25.88	19785.23
27.56	46.82	37.19	37.84	20298.52
7.94	9.33	8.64	8.67	115.40
8.11	9.39	8.75	8.81	227.09
8.21	9.54	8.88	8.98	445.36
8.45	10.00	9.23	9.43	848.04
9.00	11.12	10.06	10.47	1528.73
11.71	15.97	13.84	14.65	2184.17
15.89	24.47	20.18	21.78	2939.07
17.31	26.98	22.15	24.17	3309.25
7.56	8.03	7.80	7.83	127.76
8.08	8.58	8.33	8.36	239.15
8.11	8.65	8.38	8.41	475.44
8.28	8.82	8.55	8.59	931.51
8.41	9.08	8.75	8.79	1820.59
8.62	9.42	9.02	9.07	3526.69
9.94	11.48	10.71	10.79	5929.59
11.54	14.52	13.03	13.19	9706.59
14.88	20.43	17.66	17.92	14281.85
20.99	29.29	25.14	25.53	15042.83
23.55	34.64	29.10	29.61	17292.94
44.80	66.17	55.49	56.45	18140.95
9.27	10.96	10.12	10.16	98.47
9.76	11.52	10.64	10.71	186.76
9.92	11.56	10.74	10.86	368.16
10.15	11.96	11.06	11.29	708.40
10.68	12.76	11.72	12.18	1313.68
11.65	14.87	13.26	14.18	2256.24
16.10	22.37	19.24	21.09	3034.40
19.61	29.00	24.31	27.11	3540.68
14.17	14.98	14.58	14.64	68.31
15.32	16.11	15.72	15.78	126.73
16.20	17.13	16.67	16.73	239.10
16.79	17.68	17.24	17.31	462.20
16.91	17.84	17.38	17.46	916.57
16.45	17.54	17.00	17.11	1870.28
18.90	20.34	19.62	19.79	3233.24
21.39	24.25	22.82	23.11	5537.57
30.39	35.64	33.02	33.55	7631.02
41.63	48.67	45.15	45.93	8360.19

46.74	56.21	51.48	52.51	9751.08
80.19	98.13	89.16	91.10	11240.42
17.58	20.92	19.25	19.32	51.75
18.59	21.90	20.25	20.38	98.14
19.50	22.89	21.20	21.45	186.44
20.16	23.59	21.88	22.36	357.81
20.43	23.82	22.13	23.06	693.71
20.46	24.39	22.43	24.28	1317.75
24.47	29.88	27.18	30.90	2070.88
30.41	40.53	35.47	41.14	2333.34
32.41	43.51	37.96	45.51	2812.43

CUDA 12.0	FasterTransformer		
mem(GiB)	total avg(ms)	throughput(tokens/s)	mem(GiB)
13.26	6.82	146.63	13.57
13.34	7.39	270.64	13.70
13.50	7.39	541.27	13.95
13.82	7.56	1058.20	14.48
14.47	7.53	2124.83	15.54
15.77	7.79	4107.83	17.64
18.36	9.26	6911.45	21.85
23.57	11.03	11604.71	30.26
33.97	16.34	15667.07	47.11
44.36	21.99	17462.48	63.95
54.78			
75.57			
13.86	8.56	116.82	14.79
14.56	9.02	221.73	16.14
15.93	9.25	432.43	18.82
18.70	10.00	800.00	24.23
24.22	12.14	1317.96	35.01
35.27	16.86	1897.98	56.61
57.37			
68.42			
13.16	7.98	125.31	13.98
13.22	8.04	248.76	14.07
13.34	8.07	495.66	14.29
13.60	8.18	978.00	14.69
14.11	8.51	1880.14	15.51
15.13	9.32	3433.48	17.16
17.16	11.09	5770.96	20.48
21.23	13.95	9175.63	27.07
29.37	17.87	14325.69	40.26
37.50	24.71	15540.27	53.48
45.64	29.94	17100.87	66.66
78.19			
13.63	10.00	100.00	14.91
14.18	10.11	197.82	15.94
15.27	10.27	389.48	18.01
17.44	10.86	736.65	22.13
21.80	12.46	1284.11	30.38
30.51	16.77	1908.17	46.91
47.92			
65.33			
16.37	15.41	64.89	17.84
16.43	15.49	129.12	17.90
16.52	15.59	256.57	18.09
16.73	15.73	508.58	18.40
17.14	15.94	1003.76	19.09
17.96	17.61	1817.15	20.37
19.59	21.13	3028.87	23.03
22.84	24.22	5284.89	28.37
29.39	34.09	7509.53	38.99
35.92	45.23	8489.94	49.62

42.45	53.99	9483.24	60.24
68.59			
16.78	19.30	51.81	18.42
17.23	19.46	102.77	19.37
18.13	19.75	202.53	20.96
19.95	20.36	392.93	24.18
23.57	21.60	740.74	30.59
30.83	25.78	1241.27	43.40
45.32	35.84	1785.71	69.06
59.83			
74.33			

Ratio(%)

99.47

106.53

105.98

106.89

104.93

97.18

107.50

104.23

105.94

94.70

98.79

102.42

102.99

106.01

115.99

115.08

101.95

96.14

95.92

95.25

96.83

102.71

102.75

105.79

99.69

96.80

101.12

98.47

94.41

94.52

96.16

102.30

118.24

105.27

98.15

93.19

90.88

91.31

102.92

106.75

104.78

101.62

98.47

102.82

99.87

95.49

92.06

91.06

93.65

106.16

115.97

LLaMA model size(B)	gpu	FP16 batch	A100 40G input len	PPL.NN.LLM output len	Rev0+ prefill(ms)	
7	1	1	1	8	256	11.82
7	1	2	2	8	256	11.88
7	1	4	4	8	256	12.31
7	1	8	8	8	256	12.80
7	1	16	16	8	256	16.13
7	1	32	32	8	256	20.68
7	1	64	64	8	256	37.25
7	1	128	128	8	256	68.20
7	1	256	256	8	256	135.66
7	1	384	384	8	256	
7	1	512	512	8	256	
7	1	768	768	8	256	
7	1	1	1024	1024	1024	67.79
7	1	2	1024	1024	1024	134.00
7	1	4	1024	1024	1024	262.03
7	1	8	1024	1024	1024	516.05
7	1	16	1024	1024	1024	1020.90
7	1	32	1024	1024	1024	2035.22
7	1	64	1024	1024	1024	
7	1	80	1024	1024	1024	
13	2	1	1	8	256	13.62
13	2	2	2	8	256	13.90
13	2	4	4	8	256	14.71
13	2	8	8	8	256	16.34
13	2	16	16	8	256	20.17
13	2	32	32	8	256	28.26
13	2	64	64	8	256	41.79
13	2	128	128	8	256	75.58
13	2	256	256	8	256	136.79
13	2	384	384	8	256	205.82
13	2	512	512	8	256	
13	2	1024	1024	8	256	
13	2	1	1024	1024	1024	75.28
13	2	2	1024	1024	1024	135.19
13	2	4	1024	1024	1024	264.97
13	2	8	1024	1024	1024	526.60
13	2	16	1024	1024	1024	1057.23
13	2	32	1024	1024	1024	2069.18
13	2	64	1024	1024	1024	
13	2	96	1024	1024	1024	
65	8	1	1	8	256	
65	8	2	2	8	256	
65	8	4	4	8	256	
65	8	8	8	8	256	
65	8	16	16	8	256	
65	8	32	32	8	256	
65	8	64	64	8	256	
65	8	128	128	8	256	
65	8	256	256	8	256	
65	8	384	384	8	256	

65	8	512	8	256
65	8	1024	8	256
65	8	1	1024	1024
65	8	2	1024	1024
65	8	4	1024	1024
65	8	8	1024	1024
65	8	16	1024	1024
65	8	32	1024	1024
65	8	64	1024	1024
65	8	96	1024	1024
65	8	128	1024	1024

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/475333230122012010>