LLaMA model size(B)	gpu	FP16 batch	H800 input len	PPL.NN.LLM output len	Rev0+GEM prefill(ms)
7	1	1	8	256	6.41
7	1	2	8	256	6.50
7	1	4	8	256	6.57
7	1	8	8	256	7.10
7	1	16	8	256	7.94
7	1	32	8	256	10.41
7	1	64	8	256	16.36
7	1	128	8	256	30.03
7	1	256	8	256	58.69
7	1	384	8	256	83.42
7	1	512	8	256	110.82
7	1	768	8	256	165.19
7	1	1	1024	1024	30.86
7	1	2	1024	1024	58.33
7	1	4	1024	1024	109.13
7	1	8	1024	1024	213.48
7	1	16	1024	1024	415.95
7	1	32	1024	1024	830.35
7	1	64	1024	1024	1633.89
7	1	80	1024	1024	2078.37
13	2	1	8	256	7.60
13	2	2	8	256	7.87
13	2	4	8	256	8.09
13	2	8	8	256	9.78
13	2	16	8	256	11.10
13	2	32	8	256	13.74
13	2	64	8	256	21.33
13	2	128	8	256	40.17
13	2	256	8	256	69.08
13	2	384	8	256	99.10
13	2	512	8	256	131.19
13	2	1024	8	256	246.24
13	2	1	1024	1024	41.19
13	2	2	1024	1024	70.31
13	2	4	1024	1024	127.86
13	2	8	1024	1024	243.83
13	2	16	1024	1024	470.57
13	2	32	1024	1024	945.02
13	2	64	1024	1024	1901.01
13	2	96	1024	1024	2875.81
65	8	1	8	256	15.80
65	8	2	8	256	16.03
65	8	4	8	256	16.30
65	8	8	8	256	18.80
65	8	16	8	256	20.83
65	8	32	8	256	29.38
65	8	64	8	256	44.65
65	8	128	8	256	75.48
65	8	256	8	256	136.26
65	8	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

IM 32MiB Workspae if M >= 64

decode min(ms)	decode max(ms)	decode avg(ms)	total avg(ms)	throughput(tokens/s)
6.29	6.69	6.49	6.52	153.49
6.35	6.70	6.53	6.55	305.33
6.45	6.81	6.63	6.66	600.99
6.44	6.94	6.69	6.72	1190.88
6.51	7.06	6.79	6.82	2347.41
6.85	7.86	7.36	7.40	4326.86
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
7.71	8.98	8.35	8.38	119.40
7.72	9.12	8.42	8.48	235.93
7.84	9.19	8.52	8.62	463.95
8.14	9.74	8.94	9.15	874.46
8.64	10.80	9.72	10.13	1580.06
11.03	15.38	13.21	14.02	2283.12
15.89	24.47	20.18	21.78	2939.07
17.31	26.98	22.15	24.17	3309.25
7.54	7.97	7.76	7.78	128.46
7.62	8.02	7.82	7.85	254.75
7.63	8.02	7.83	7.86	509.13
7.73	8.23	7.98	8.02	997.73
7.81	8.50	8.16	8.20	1951.61
8.14	8.99	8.57	8.62	3712.87
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
9.30	10.92	10.11	10.15	98.52
9.23	10.95	10.09	10.16	196.88
9.37	11.01	10.19	10.31	387.79
9.61	11.49	10.55	10.79	741.56
10.04	12.23	11.14	11.59	1379.96
11.21	14.35	12.78	13.70	2335.28
16.10	22.37	19.24	21.09	3034.40
19.61	29.00	24.31	27.11	3540.68
14.60	15.52	15.06	15.12	66.13
15.48	16.40	15.94	16.00	124.98
15.62	16.55	16.09	16.15	247.70
15.82	16.67	16.25	16.32	490.24
15.96	16.90	16.43	16.51	969.03
16.12	17.16	16.64	16.75	1909.90
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
17.98	21.32	19.65	19.72	50.70
18.84	22.20	20.52	20.65	96.84
18.99	22.34	20.67	20.92	191.17
19.24	22.62	20.93	21.41	373.60
19.52	22.95	21.24	22.17	721.55
20.12	24.00	22.06	23.92	1337.86
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

	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(%)
104.68
112.82
111.03
112.54
110.48
105.33
107.50
104.23
105.94
94.70
102.21
106.41
107.29
109.31
119.89
120.29
102.51
102.41
102.72
102.02
103.80
108.14
102.75
105.79
99.69
96.80
101.12
98.52
99.52
99.57
100.67
107.46
122.38
101.91
96.80
96.54
96.39
96.54
105.10
106.75
104.78
101.62
98.47

102.82

97.85
94.22
94.39
95.08
97.41
107.78
115.97

LLaMA model size(B)	gpu	FP16 batch	A100 40G input len	PPL.NN.LLM output len	prefill(ms)
7	1	1	8	256	11.90
7	1	2	8	256	11.87
7	1	4	8	256	12.17
7	1	8	8	256	12.78
7	1	16	8	256	14.78
7	1	32	8	256	20.38
7	1	64	8	256	41.75
7	1	128	8	256	70.64
7	1	256	8	256	134.76
7	1	384	8	256	
7	1	512	8	256	
7	1	768	8	256	
7	1	1	1024	1024	69.39
7	1	2	1024	1024	131.41
7	1	4	1024	1024	248.09
7	1	8	1024	1024	493.11
7	1	16	1024	1024	1165.47
7	1	32	1024	1024	1969.27
7	1	64	1024	1024	
7	1	80	1024	1024	
13	2	1	8	256	13.39
13	2	2	8	256	13.58
13	2	4	8	256	14.23
13	2	8	8	256	15.59
13	2	16	8	256	20.29
13	2	32	8	256	27.50
13	2	64	8	256	45.40
13	2	128	8	256	71.53
13	2	256	8	256	135.95
13	2	384	8	256	203.73
13	2	512	8	256	
13	2	1024	8	256	
13	2	1	1024	1024	71.19
13	2	2	1024	1024	134.29
13	2	4	1024	1024	260.67
13	2	8	1024	1024	516.58
13	2	16	1024	1024	1030.02
13	2	32	1024	1024	2031.51
13	2	64	1024	1024	
13	2	96	1024	1024	
65	8	1	8	256	
65	8	2	8	256	
65	8	4	8	256	
65	8	8	8	256	
65	8	16	8	256	
65	8	32	8	256	
65	8	64	8	256	
65	8	128	8	256	
65	8	256	8	256	
65	8	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