Processing: 15it [06:15, 24.923/it]
2025-02-22 09:58:17,379 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	373.748
Number of concurrency	1
Total requests	15
Succeed requests	15
Failed requests	0
Throughput(average tokens/s)	44.589
Average QPS	0.04
Average latency (s)	24.909
Average time to first token (s)	0.049
Average time per output token (s)	0.02243
Average input tokens per request	11.0
Average output tokens per request	1111.0
Average package latency (s)	0.022
Average package per request	1111.0
Expected number of requests	15
Result DB path	./outputs/20250222_095203/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 09:58:17,389 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.0438	0.022	24.8556	11	1111	44.42
25%	0.0487	0.0222	24.8781	11	1111	44.5672
50%	0.05	0.0224	24.9055	11	1111	44.6085
66%	0.0504	0.0225	24.9264	11	1111	44.6533
75%	0.0509	0.0226	24.9286	11	1111	44.6578
80%	0.0519	0.0226	24.9346	11	1111	44.6705
90%	0.0527	0.0227	25.0112	11	1111	44.6982
95%	0.0536	0.0229	25.0358	11	1111	44.777
98%	0.0536	0.0231	25.0358	11	1111	44.777
99%	0.0536	0.0233	25.0358	11	1111	44.777

(base) root@gpu-1065fbd5b32ab9e4005ba-1-n74vv52lqmui:~/data/evalscope#

4090单卡1个并发

4090单卡2个并发

```
Processing: 15it [02:41, 10.79s/it]
2025-02-22 10:03:09,889 - evalscope - INFO -
Benchmarking summary:
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Key                             | Value                                                                                               |
+=================================+=====================================================================================================+
| Time taken for tests (s)        | 161.8                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Number of concurrency           | 2                                                                                                   |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Total requests                  | 15                                                                                                  |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Succeed requests                | 15                                                                                                  |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Failed requests                 | 0                                                                                                   |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Throughput(average tokens/s)    | 81.452                                                                                              |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average QPS                     | 0.093                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average latency (s)             | 19.909                                                                                              |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average time to first token (s) | 0.055                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average time per output token (s) | 0.01228                                                                                           |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average input tokens per request | 11.0                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average output tokens per request | 878.6                                                                                             |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average package latency (s)     | 0.023                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Average package per request     | 878.6                                                                                               |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Expected number of requests     | 15                                                                                                  |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
| Result DB path                  | ./outputs/20250222_100027/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db                 |
+---------------------------------+-----------------------------------------------------------------------------------------------------+
2025-02-22 10:03:09,897 - evalscope - INFO -
Percentile results:
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
| Percentile | TTFT (s) | TPOT (s) | Latency (s) | Input tokens | Output tokens | Throughput(tokens/s) |
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
|    10%     |  0.0512  |  0.0222  |   19.5084   |      11      |      862      |        43.7789          |
|    25%     |  0.0513  |  0.0224  |   19.5128   |      11      |      862      |        44.1068          |
|    50%     |  0.0525  |  0.0226  |   19.5421   |      11      |      862      |        44.1228          |
|    66%     |  0.0559  |  0.0228  |   19.5433   |      11      |      862      |        44.1611          |
|    75%     |  0.0575  |  0.0228  |   19.548    |      11      |      862      |        44.1765          |
|    80%     |  0.0584  |  0.0229  |   19.6898   |      11      |      862      |        44.1862          |
|    90%     |  0.0597  |  0.023   |   19.6961   |      11      |      862      |        44.1865          |
|    95%     |  0.0725  |  0.0231  |   24.909    |      11      |     1111      |        44.6023          |
|    98%     |  0.0725  |  0.0234  |   24.909    |      11      |     1111      |        44.6023          |
|    99%     |  0.0725  |  0.0236  |   24.909    |      11      |     1111      |        44.6023          |
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
(base) root@gpu-1065fbd5b32ab9e4005ba-1-n74vv52lqmui:~/data/evalscope#
```

4090单卡5个并发

Processing: 15it [01:13, 4.92s/it]
2025-02-22 10:06:23,870 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	73.831
Number of concurrency	5
Total requests	15
Succeed requests	15
Failed requests	0
Throughput(average tokens/s)	212.31
Average QPS	0.203
Average latency (s)	24.6
Average time to first token (s)	0.06
Average time per output token (s)	0.00471
Average input tokens per request	11.0
Average output tokens per request	1045.0
Average package latency (s)	0.023
Average package per request	1045.0
Expected number of requests	15
Result DB path	./outputs/20250222_100509/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 10:06:23,879 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.0517	0.0228	24.5754	11	1045	42.4166
25%	0.052	0.0231	24.5763	11	1045	42.4273
50%	0.0589	0.0235	24.5943	11	1045	42.4896
66%	0.0594	0.0237	24.5945	11	1045	42.5024
75%	0.0717	0.0239	24.6303	11	1045	42.5207
80%	0.0722	0.024	24.6304	11	1045	42.5212
90%	0.0725	0.0242	24.6366	11	1045	42.5222
95%	0.0727	0.0244	24.6366	11	1045	42.5223
98%	0.0727	0.0246	24.6366	11	1045	42.5223
99%	0.0727	0.0247	24.6366	11	1045	42.5223

4090单卡10个并发

```
Processing: 50it [03:32,  4.26s/it]
2025-02-22 10:14:10,799 - evalscope - INFO -
Benchmarking summary:
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Key                               | Value                                                                                               |
+===================================+=====================================================================================================+
| Time taken for tests (s)          | 212.771                                                                                             |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Number of concurrency             | 10                                                                                                  |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Total requests                    | 50                                                                                                  |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Succeed requests                  | 50                                                                                                  |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Failed requests                   | 0                                                                                                   |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Throughput(average tokens/s)      | 286.65                                                                                              |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average QPS                       | 0.235                                                                                               |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average latency (s)               | 39.67                                                                                               |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average time to first token (s)   | 1.323                                                                                               |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average time per output token (s) | 0.00349                                                                                             |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average input tokens per request  | 11.0                                                                                                |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average output tokens per request | 1219.82                                                                                             |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average package latency (s)       | 0.031                                                                                               |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Average package per request       | 1218.6                                                                                              |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Expected number of requests       | 50                                                                                                  |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
| Result DB path                    | ./outputs/20250222_101037/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db                 |
+-----------------------------------+-----------------------------------------------------------------------------------------------------+
2025-02-22 10:14:10,830 - evalscope - INFO -
Percentile results:
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
| Percentile | TTFT (s) | TPOT (s) | Latency (s) | Input tokens | Output tokens | Throughput(tokens/s) |
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
|    10%     |  0.0571  |  0.024   |   28.1172   |      11      |      937      |        24.2599          |
|    25%     |  0.0609  |  0.0244  |   33.5314   |      11      |     1048      |        27.6156          |
|    50%     |  0.161   |  0.0247  |   38.0579   |      11      |     1219      |        31.4326          |
|    66%     |  0.3317  |  0.0249  |   42.5026   |      11      |     1346      |        35.0606          |
|    75%     |  1.0068  |  0.0251  |   46.0029   |      11      |     1380      |        36.1213          |
|    80%     |  1.1692  |  0.0252  |   46.7057   |      11      |     1422      |        36.1969          |
|    90%     |  7.4921  |  0.0254  |   51.1389   |      11      |     1558      |        38.3841          |
|    95%     |  9.0909  |  0.0257  |   62.6667   |      11      |     1707      |        41.0967          |
|    98%     |  9.0991  |  0.026   |   65.7853   |      11      |     1948      |        41.0968          |
|    99%     |  9.0991  |  0.0268  |   65.7853   |      11      |     1948      |        41.0968          |
+------------+----------+----------+-------------+--------------+---------------+-------------------------+
(base) root@gpu-1065fbd5b32ab9e4005ba-1-n74vv52lgmui:~/data/evalscope#
```

4090单卡20个并发

Processing: 100it [08:11, 4.913s/it]
2025-02-22 10:23:19,367 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	491.216
Number of concurrency	20
Total requests	100
Succeed requests	100
Failed requests	0
Throughput(average tokens/s)	273.946
Average QPS	0.204
Average latency (s)	89.549
Average time to first token (s)	9.521
Average time per output token (s)	0.00365
Average input tokens per request	11.0
Average output tokens per request	1345.67
Average package latency (s)	0.06
Average package per request	1341.02
Expected number of requests	100
Result DB path	./outputs/20250222_101507/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 10:23:19,439 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.171	0.0241	68.3796	11	981	11.3187
25%	0.1869	0.0244	78.2333	11	1143	12.6893
50%	3.5876	0.0248	87.5852	11	1353	14.3191
66%	11.9239	0.0251	97.4601	11	1459	16.5233
75%	15.7769	0.0255	106.4783	11	1539	17.1401
80%	19.2604	0.0258	108.2781	11	1649	17.9611
90%	32.3422	0.0278	119.3135	11	1728	21.1334
95%	37.1891	0.0282	125.2375	11	1863	33.3593
98%	40.8693	0.0287	127.7291	11	2048	39.056
99%	40.8791	0.0578	136.4465	11	2048	39.0562

4090单卡30个并发

Processing: 90it [07:24, 4.94s/it]
2025-02-22 11:37:14,436 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	444.973
Number of concurrency	30
Total requests	90
Succeed requests	90
Failed requests	0
Throughput(average tokens/s)	273.452
Average QPS	0.202
Average latency (s)	126.781
Average time to first token (s)	14.514
Average time per output token (s)	0.00366
Average input tokens per request	11.0
Average output tokens per request	1351.989
Average package latency (s)	0.083
Average package per request	1346.922
Expected number of requests	90
Result DB path	./outputs/20250222_112949/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 11:37:14,496 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.2192	0.0241	55.6977	11	895	6.4573
25%	0.2213	0.0244	119.2692	11	1122	8.3005
50%	5.0742	0.0248	142.4192	11	1375	10.1789
66%	17.95	0.0251	147.2518	11	1480	11.3701
75%	25.1508	0.0255	152.7669	11	1570	12.2877
80%	27.9337	0.026	154.6278	11	1630	13.8774
90%	45.2813	0.0279	160.6616	11	1814	28.3486
95%	53.6729	0.0288	165.1017	11	1932	38.7107
98%	58.1855	0.03	171.0384	11	2048	38.7117
99%	65.5253	0.0565	174.5434	11	2048	38.712

(base) root@gpu-1065fbd5b32ab9e400f5ba-1-n74vw52lgmui:~/data/evalscope#

4090卡40个并发

}
Processing: 80it [07:11, 5.39s/it]
2025-02-22 11:58:32,723 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	431.472
Number of concurrency	40
Total requests	80
Succeed requests	65
Failed requests	15
Throughput(average tokens/s)	218.872
Average QPS	0.151
Average latency (s)	150.218
Average time to first token (s)	32.258
Average time per output token (s)	0.00457
Average input tokens per request	11.0
Average output tokens per request	1452.877
Average package latency (s)	0.081
Average package per request	1448.462
Expected number of requests	80
Result DB path	./outputs/20250222_115120/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 11:58:32,770 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.2642	0.0241	46.7892	11	994	5.5723
25%	0.2665	0.0244	95.3286	11	1170	7.5791
50%	1.1436	0.0247	176.0987	11	1461	8.9775
66%	31.7241	0.0251	185.8983	11	1602	11.5496
75%	61.4038	0.0254	194.4921	11	1707	13.9196
80%	85.9034	0.0258	203.6297	11	1750	17.5505
90%	98.5896	0.0285	220.3502	11	2027	33.3364
95%	116.8934	0.0304	235.3202	11	2048	37.8863
98%	116.8936	0.0321	238.0755	11	2048	37.8874
99%	116.9002	0.0334	238.1783	11	2048	37.8877

(base) root@gpu-1065fbd5b32ab9e4005ba-1-n74vv52lgmui:~/data/evalscope#

4090单卡50个并发

Processing: 500it [37:06, 4.45s/it]
2025-02-22 11:27:15,254 - evalscope - INFO -
Benchmarking summary:

Key	Value
Time taken for tests (s)	2226.191
Number of concurrency	50
Total requests	500
Succeed requests	395
Failed requests	105
Throughput(average tokens/s)	230.566
Average QPS	0.177
Average latency (s)	211.593
Average time to first token (s)	40.089
Average time per output token (s)	0.00434
Average input tokens per request	11.0
Average output tokens per request	1299.451
Average package latency (s)	0.133
Average package per request	1292.909
Expected number of requests	500
Result DB path	./outputs/20250222_105008/DeepSeek-R1-Distill-Qwen-32B-Int4-W4A16/benchmark_data.db

2025-02-22 11:27:15,534 - evalscope - INFO -
Percentile results:

Percentile	TTFT (s)	TPOT (s)	Latency (s)	Input tokens	Output tokens	Throughput(tokens/s)
10%	0.5488	0.0241	169.9692	11	891	4.2021
25%	12.6779	0.0244	210.6571	11	1081	5.0045
50%	38.0821	0.0248	222.4627	11	1291	5.878
66%	53.4952	0.0252	227.5406	11	1411	6.4675
75%	63.0625	0.0255	231.1495	11	1499	6.883
80%	71.7708	0.0257	232.7389	11	1555	7.2044
90%	85.4972	0.0286	235.9382	11	1720	8.6627
95%	95.9236	0.0322	237.8129	11	1840	11.4826
98%	109.2154	0.0393	239.2548	11	2028	28.5417
99%	110.5776	0.1259	239.9838	11	2048	35.7509