模型效果评测

Refresh

中文评测

模型类型	架构	参数大小	模型名称	厂商	逻辑推理	表格问答	复杂指令	数学能力	总分	加权总分	测试方式
开源模型	Dense	340B	Nemotron-4-340B	英伟达	26	64	25	19	47.14	42.71	API
开源模型	Dense	115B	TeleChat2-115B	中国电信	28	61	14	40	51.30	37.07	本地部署
开源模型	Dense	72B	Qwen2.5-72B-Instruct	阿里巴巴	43	75	33	50	73.57	64.17	本地部署
开源模型	Dense	67B	deepseek-67B	幻方-深度求索	26	64	21	16	44.23	37.84	本地部署
开源模型	Dense	32B	Qwen2.5-32B-Instruct	阿里巴巴	44	74	32	49	72.87	63.06	本地部署
开源模型	Dense	20B	internlm2-5-20b-chat	上海人工智能实验室	26	60	17	26	45.60	35.96	本地部署
开源模型	Dense	14B	Qwen2.5-14B-Instruct	阿里巴巴	36	77	24	49	67.05	53.07	本地部署
开源模型	Dense	14B	Orion-14B-Chat	猎户星空	9	46	8	3	21.25	15.26	本地部署
开源模型	Dense	13B	baichuan-13B	百川智能	8	54	3	14	25.59	13.39	本地部署
开源模型	Dense	9B	glm4-9B-chat	智谱华章	24	55	16	37	47.53	36.80	本地部署
开源模型	Dense	8B	internlm2-chat-1_8b	上海人工智能实验室	6	23	2	2	10.63	6.09	本地部署
开源模型	Dense	7B	Qwen2.5-7B-Instruct	阿里巴巴	23	52	11	23	38.36	27.57	本地部署
开源模型	Dense	7B	baichuan-7B	百川智能	11	49	5	11	25.01	15.00	本地部署
开源模型	Dense	4B	MiniCPM3-4B	面壁智能	16	50	8	30	36.54	24.32	本地部署
开源模型	Dense	3.8B	Phi-3.5-mini-instruct	微软	18	42	7	19	30.13	20.30	本地部署
开源模型	Dense	3B	Qwen2.5-3B-Instruct	阿里巴巴	21	49	5	39	40.85	24.75	本地部署
开源模型	Dense	3B	llama3.2-3B-instruct	meta	15	42	5	8	23.46	14.66	本地部署
开源模型	Dense	2B	MiniCPM-2B-dpo-bf16	面壁智能	8	32	6	10	18.99	13.54	本地部署
开源模型	Dense	1.5B	Qwen2.5-1.5B-Instruct	阿里巴巴	19	33	4	14	24.69	15.42	本地部署
开源模型	Dense	1B	llama3.2-1B-instruct	META	9	22	2	6	13.26	7.80	本地部署
开源模型	Dense	0.5B	Qwen2.5-0.5B-Instruct	阿里巴巴	11	24	1	11	16.32	8.74	本地部署
闭源模型	Dense		gpt-4o-20240513	OpenAI	38	77	35	47	71.63	64.35	API
闭源模型	Dense		gpt-4-1106-preview	OpenAI	37	81	32	42	69.00	60.08	API
闭源模型	Dense		gpt-4o-mini-2024-07-18	OpenAI	31	62	25	48	60.75	51.26	API
闭源模型	Dense		step-2-16k	阶跃星辰	38	76	30	45	68.43	58.62	API

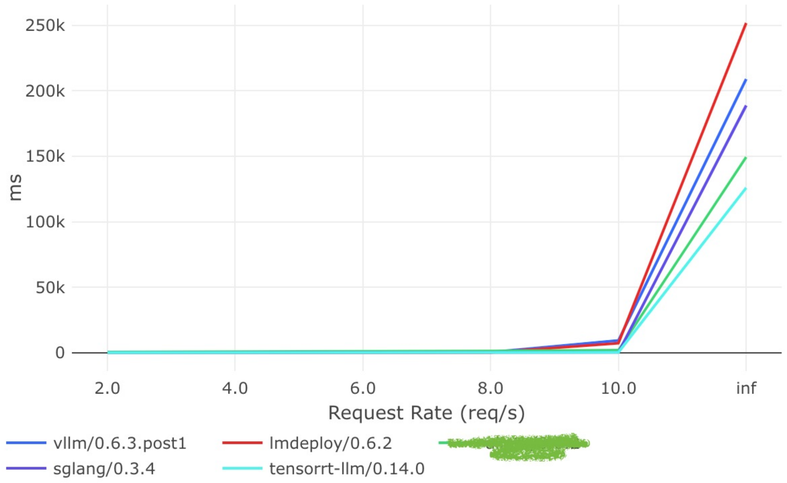
查询日期: Last 12 months | 模型: Qwen2-72B-Instruct | 量化方法: default | 数据集: sharegpt | 芯片: [illegible] | 引擎/版本: vllm/0.6.3... × lmdeploy/0... × sglang/0.3... × +2 more

竞品对比

- 不同引擎 or 版本按照所筛选的时间范围内最新测试记录进行对比
- 目前仅 [illegible] sharegpt数据集 拥有足够竞品模型数据可进行横向对比
- 大盘最底部包含筛选对比的数据明细 当 `dataset` 为 `'sharegpt'` 且 `request_num = 4000` 时，不同模型对 `request_rate_float` 限制如下：
  - [illegible]: `[4, 8, 12, 16, 20, 24, inf]`
  - [illegible]: `[2, 4, 6, 8, 10, inf]`

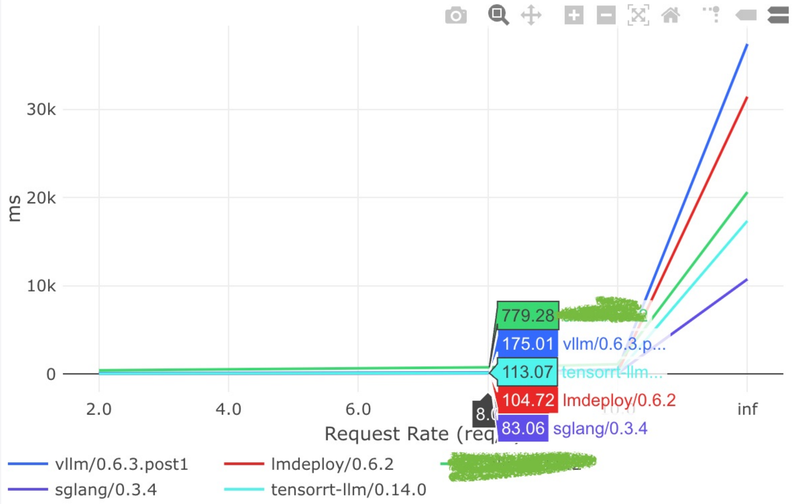
（`inf` 表示无穷大）

TTFT avg – 不同request_rate下竞品对比



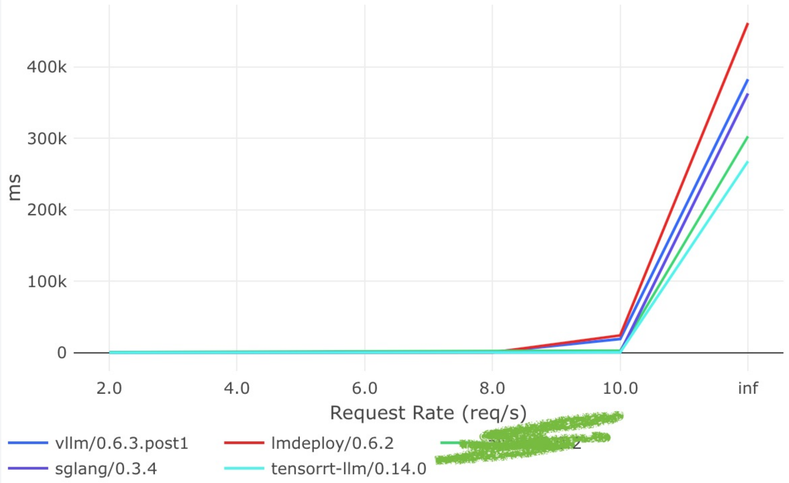
11 minutes ago

TTFT p10 – 不同request_rate下竞品对比



11 minutes ago

TTFT p90 – 不同request_rate下竞品对比



11 minutes ago

TPOT avg – 不同request_rate下竞品对比

TPOT p10 – 不同request_rate下竞品对比

TPOT p90 – 不同request_rate下竞品对比

Dashboards
Queries
Alerts
Create
Help
Settings

sglang/0.3.4 tensorrt-llm/0.14.0

12 minutes ago

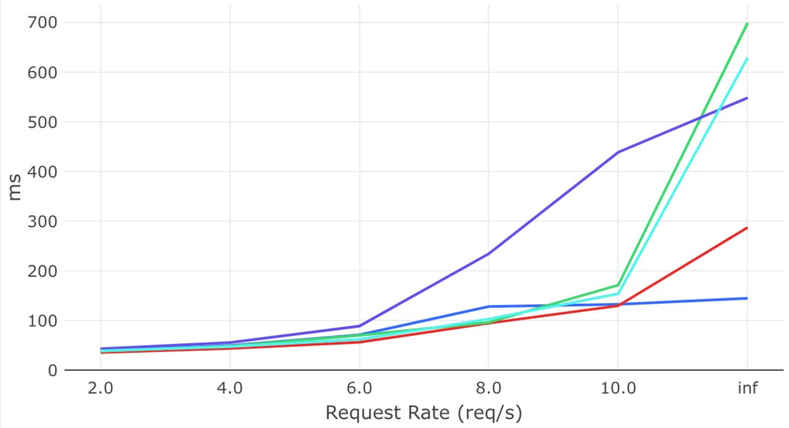
sglang/0.3.4 tensorrt-llm/0.14.0

11 minutes ago

sglang/0.3.4 tensorrt-llm/0.14.0

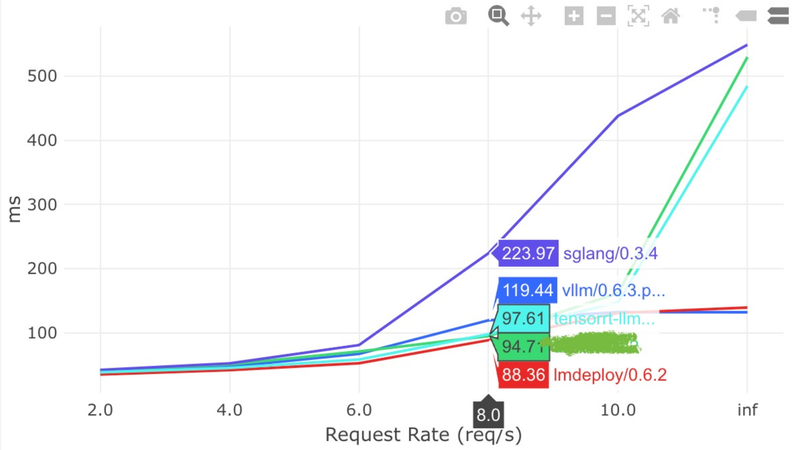
12 minutes ago

TPOT avg – 不同request_rate下竞品对比



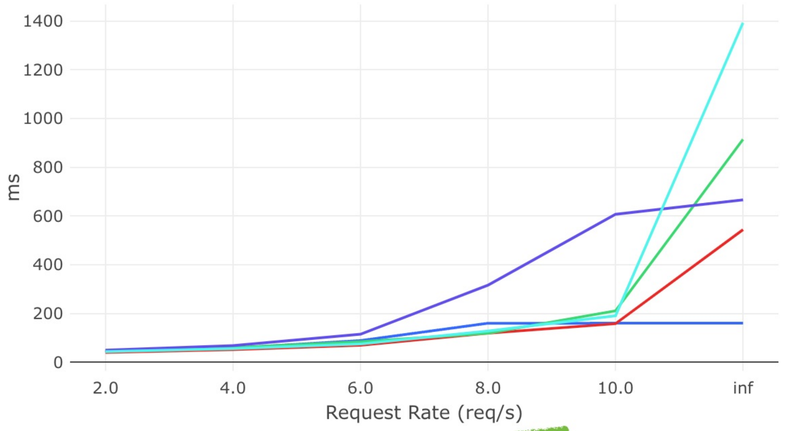
12 minutes ago

TPOT p10 – 不同request_rate下竞品对比



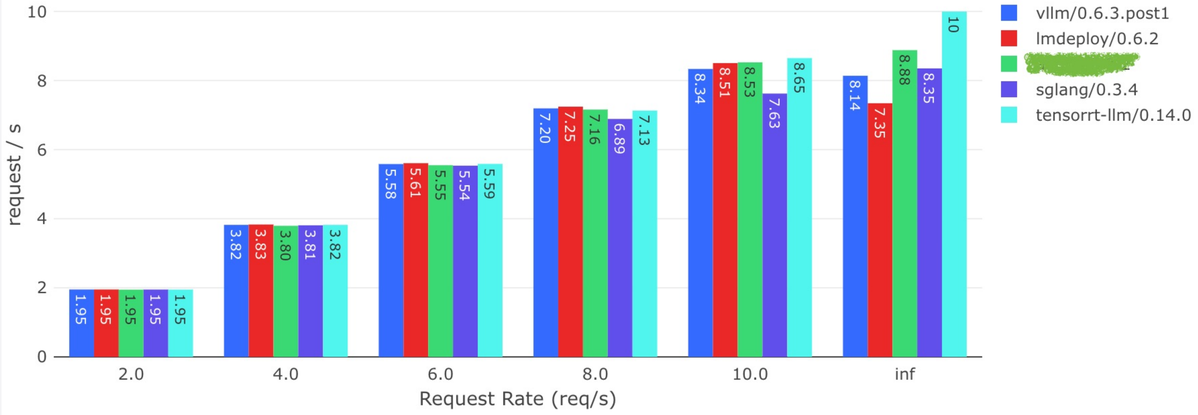
12 minutes ago

TPOT p90 – 不同request_rate下竞品对比



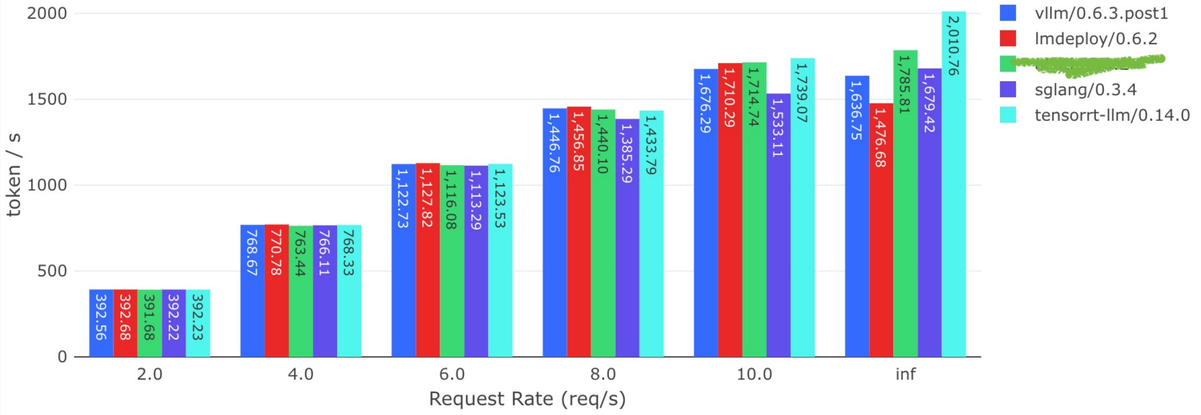
12 minutes ago

平均request / s – 不同request_rate下竞品对比



12 minutes ago

平均输出 tokens / s – 不同request_rate下竞品对比



11 minutes ago

不同request_rate下竞品对比

Show Results Only

评测平台ClickHouse

Search schema...

- INFORMATION_SCHEMA.referential_constraints
- INFORMATION_SCHEMA.schemata
- INFORMATION_SCHEMA.statistics
- INFORMATION_SCHEMA.tables
- INFORMATION_SCHEMA.views
- chip_eval.hardware_info
- chip_eval.static_inference_performance_full
- chip_eval.static_inference_performance_latest
- chip_eval.static_inference_performance_model
- chip_eval.static_inference_platform_provider_
- chip_eval.static_inference_platform_provider_4090
- chip_eval.static_inference_platform_provider_template
- evaluation.hardware_info
- evaluation.inference_evaluation_en
- evaluation.inference_evaluation_statistic
- evaluation.inference_evaluation_zh
- evaluation.serving_inference_performance
- evaluation.static_inference_performance
- evaluation.static_inference_performance_full
- evaluation.static_inference_performance_model
- information_schema.COLUMNS
- information_schema.KEY_COLUMN_USAGE
- information_schema.REFERENTIAL_CONSTRAINTS
- information_schema.SCHEMATA
- information_schema.STATISTICS
- information_schema.TABLES
- information_schema.VIEWS
- information_schema.columns
- information_schema.key_column_usage
- information_schema.referential_constraints
- information_schema.schemata
- information_schema.statistics
- information_schema.tables
- information_schema.views

Add description

chenyonghua created a year ago

chenyonghua updated 12 minutes ago

Refresh Schedule Every hour

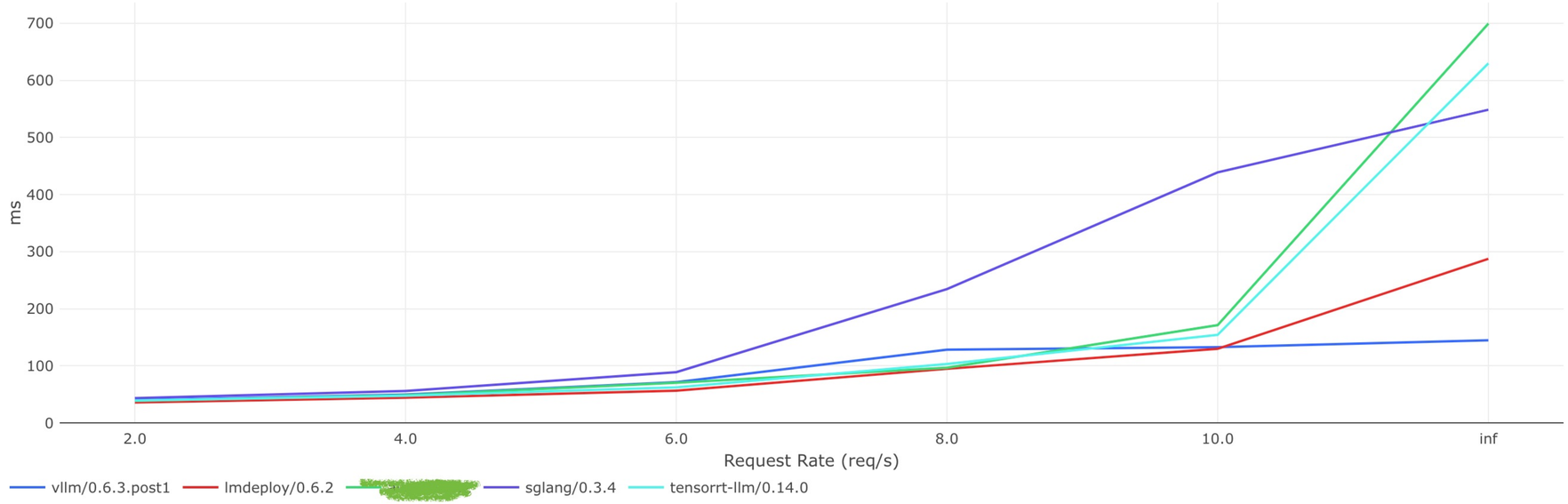
```sql
SELECT
  *
FROM
  (
    SELECT
      row_number() OVER (
        PARTITION BY engine_id,
        request_rate
        ORDER BY
          timestamp DESC
      ) AS rn,
      *
    FROM
      (
        SELECT
          *,
          dense_rank() OVER (
```

{{ }} LIMIT 1000 Save Execute

查询日期 | 模型 | 量化方法 | 数据集 | 芯片 | 引擎/版本

Last 12 months | Qwen2-72B-Instruct | default | sharegpt | [redacted] | vllm/0.6.3... | lmdeploy/0... | sglang/0.3... | +2 more

数据明细 | TTFT avg | TTFT p10 | TTFT p90 | TPOT avg | TPOT p10 | TPOT p90 | 请求成功率 | 平均request / s | 平均输出 tokens / s | Add Visualization



Edit Visualization 30 rows 0 seconds runtime Refreshed 12 minutes ago

Dashboards Queries Alerts Create Help Settings

Serving推理性能数据大盘

Refresh

日期: Last 12 months

引擎/版本: [redacted] × [redacted] × vllm-nvidi... ×

芯片: [redacted]

模型: Qwen2.5-72... × Qwen2.5-32... × Qwen2.5-14... × +1 more

数据集: maas_qwen2... ×

量化方法: default × awq × fp8 ×

数据明细 – Serving推理性能数据总表（含MaaS数据）

测试时间	芯片规格	引擎/版本	模型	量化方法	数据集	请求总量	并发数	请求速率	请求成功率	输入总token量	输出总token量	TTFT avg	TPOT avg
2024-11-26 15:19:37	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	512	串行	100.00%	424,254	82,890	27,788.16	263.65
2024-11-26 15:17:18	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	256	串行	100.00%	424,254	82,835	9,740.91	197.68
2024-11-26 15:15:02	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	128	串行	100.00%	424,254	82,808	2,542.85	136.91
2024-11-26 15:12:36	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	64	串行	100.00%	424,254	82,783	1,053.86	87.69
2024-11-26 15:09:47	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	32	串行	100.00%	424,254	82,798	587.81	60.54
2024-11-26 15:06:11	[redacted]	[redacted]	Qwen2.5-32B-Instruct	default	maas_qwen2_in424_out85_num1000	1,000	16	串行	100.00%	424,254	82,802	415.01	44.57

< 1 2 3 4 >

18 minutes ago

引擎模型覆盖度 – 引擎评测模型覆盖度

公司/组织	推理框架	Meta-Llama-3-70B-Instruct	Meta-Llama-3-8B-Instruct	Qwen2-72B-Instruct	Qwen2-7B-Instruct	Qwen1.5-72B-Chat	Qwen1.5-7B-Chat	Qwen2-72B-Instruct-awq	Qwen1.5-72B-Chat-awq	Meta-Llama-3-7
[redacted]	[redacted]	✅	✅	✅	✅	✅		✅	✅	
[redacted]	[redacted]	✅	✅	✅	✅	✅	✅	✅	✅	
腾讯	KsanaLLM						✅			
ModelTC	lightLLM	✅	✅							
上海人工智能实验室	LMDeploy	✅	✅	✅	✅	✅	✅	✅	✅	
Vectorch AI	ScaleLLM	✅	✅	✅		✅		✅	✅	
LMSYS Org	SGLang	✅	✅	✅	✅	✅		✅		
英伟达	TensorRT-LLM	✅	✅	✅	✅	✅			✅	
Hugging Face	TGI	✅	✅			✅				
UC Berkeley	vllm	✅	✅	✅	✅	✅	✅	✅	✅	

just now

Dashboards
Queries
Alerts
Create
Help
Settings