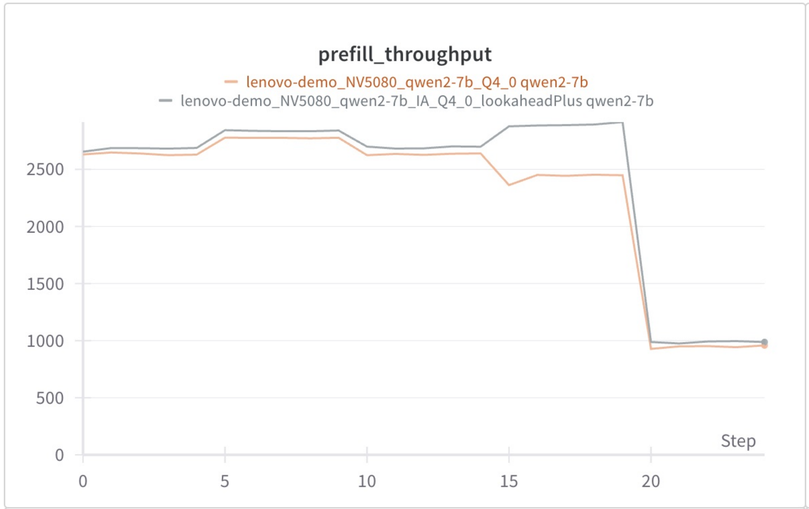
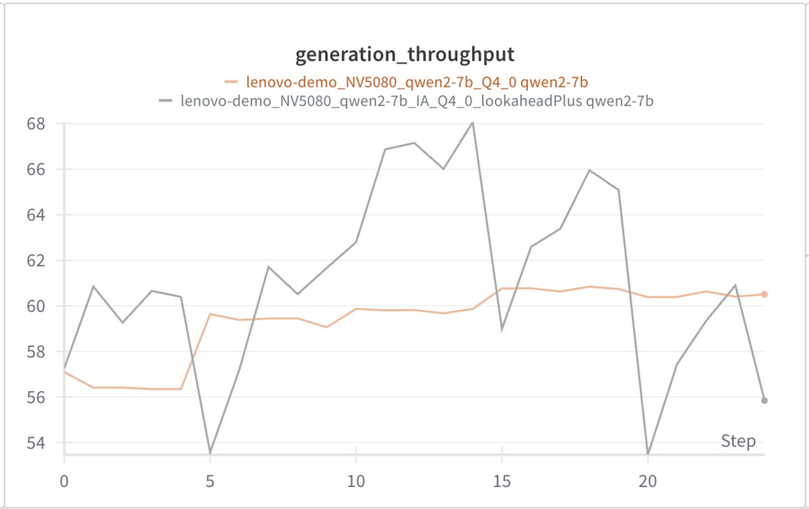
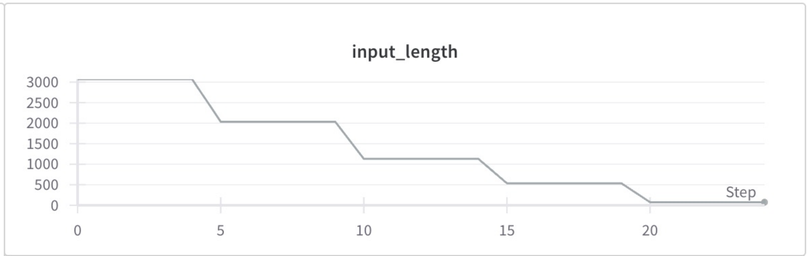
1. 从7～8b的模型来看，开启IA_Q4_0和投机后对与GPU的功耗、显存占用大致都为base的85%～90%左右;

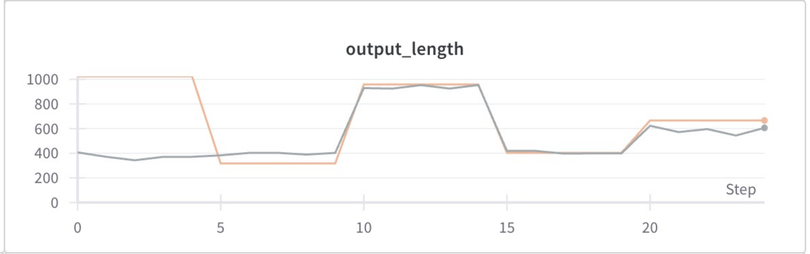
指标看板

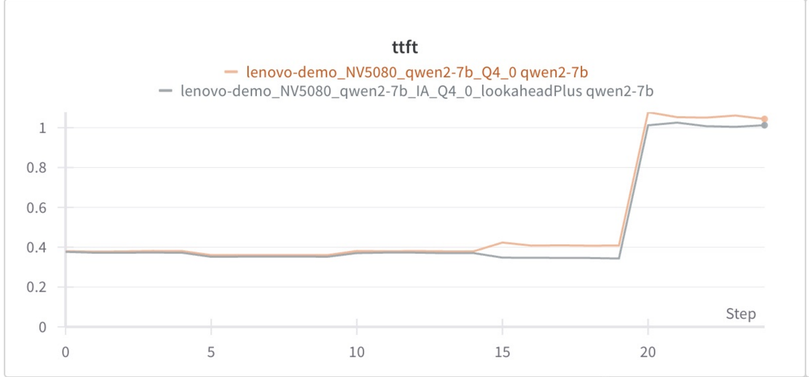
💡 下方表格勾选指定tab筛选模型（默认勾选ds-2-7b）

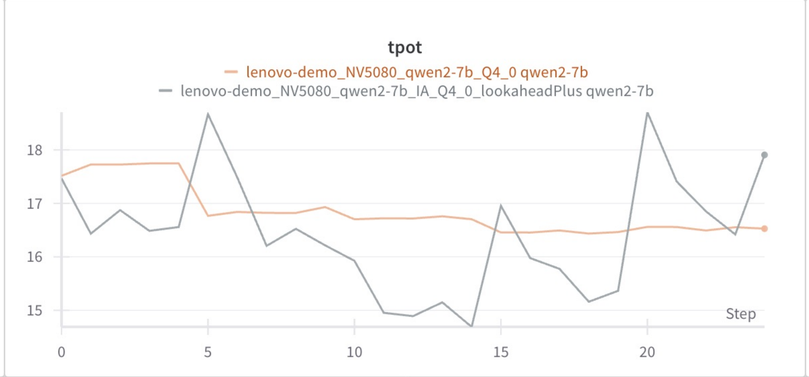


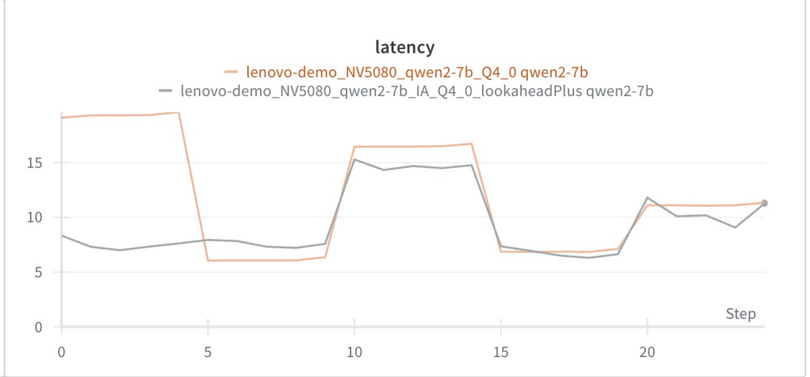


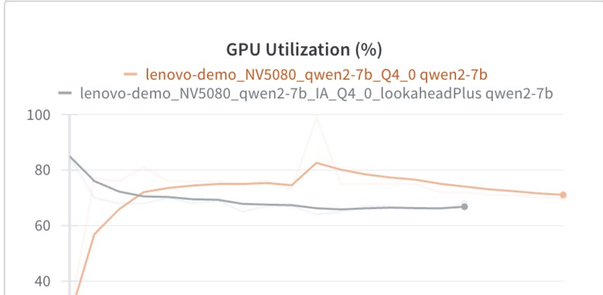


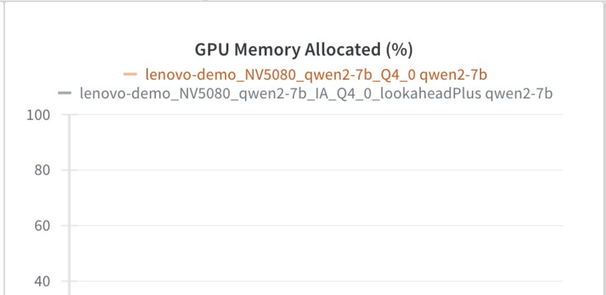


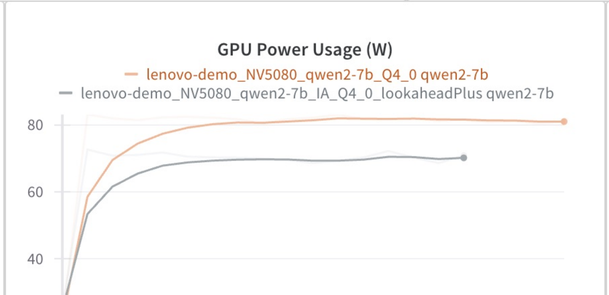


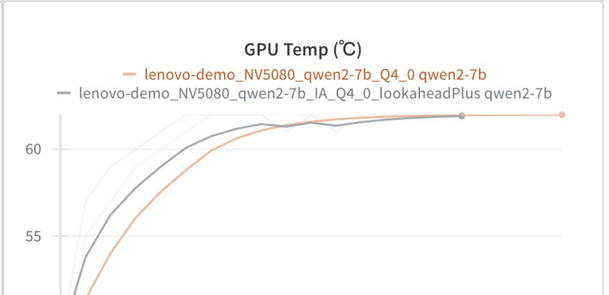












Chenyonghua's run workspace Personal workspace Autosaved just now

kimi-k2-instruct_default_[redacted]

Search panels with regex Create report

Tables 2 Add panel

```
runs.summary["benchmark_results"]
```

	芯片规格	模型	软件框架	集群规格	数据集	请求总量	请求速率	请求成功率(%	审核命中	并发数	输入总token量	输出总token量	思维链总token量	总耗时(s)	总请求吞吐(re	总输入吞吐(to	总输出吞吐(to	TTFT_avg
1	[redacted]	kimi-k2-instruct-	[redacted]	[redacted]	random	10	串行	100.0	0	1	5583	1369	0	42.32	0.24	131.92	32.35	405.71
2	[redacted]	kimi-k2-instruct-	[redacted]	[redacted]	random	10	串行	100.0	0	2	5583	1322	0	22.82	0.44	244.66	57.93	445.84
3	[redacted]	kimi-k2-instruct-	[redacted]	[redacted]	random	10	串行	100.0	0	4	5583	1363	0	14.18	0.71	393.79	96.14	530.99
4	[redacted]	kimi-k2-instruct-	[redacted]	[redacted]	random	10	串行	100.0	0	8	5583	1342	0	9.47	1.06	589.56	141.71	622.53

1 - 4 of 4 Export as CSV Columns... Reset table

Lengths Distribution

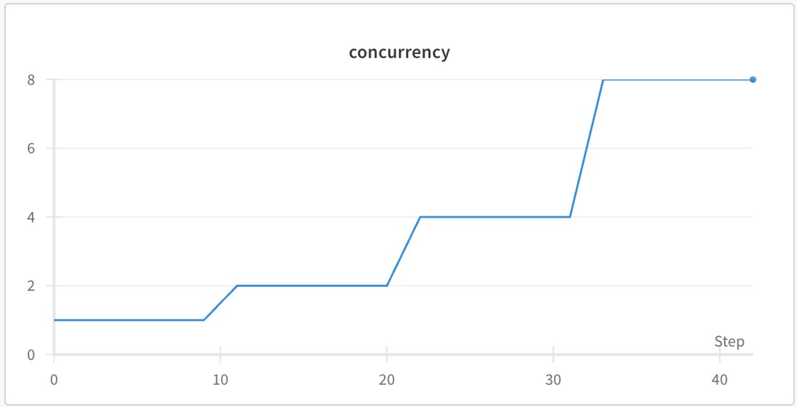
Search 1-6 of 6

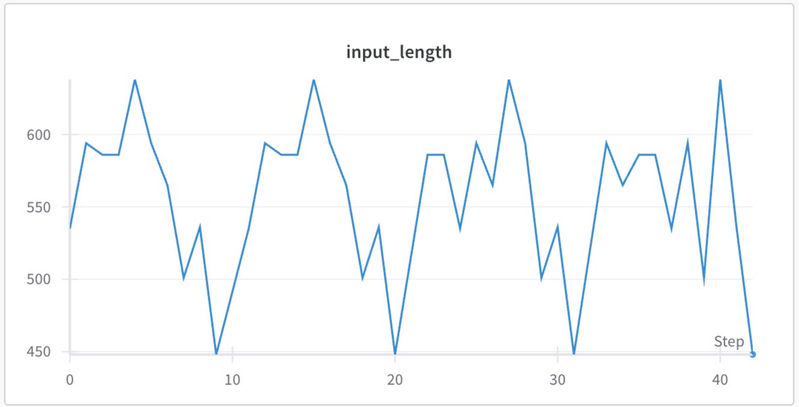
Scenario	Percentile	Prompt Input	Actual Output	Reasoning
random|Req:10|Rate:-1.0|Con:8	0%	448	106	0
random|Req:10|Rate:-1.0|Con:8	25%	535.25	128.25	0
random|Req:10|Rate:-1.0|Con:8	50%	575.5	136.5	0
random|Req:10|Rate:-1.0|Con:8	75%	592	144.5	0
random|Req:10|Rate:-1.0|Con:8	100%	638	151	0
random|Req:10|Rate:-1.0|Con:8	Avg	558.3	134.2	0

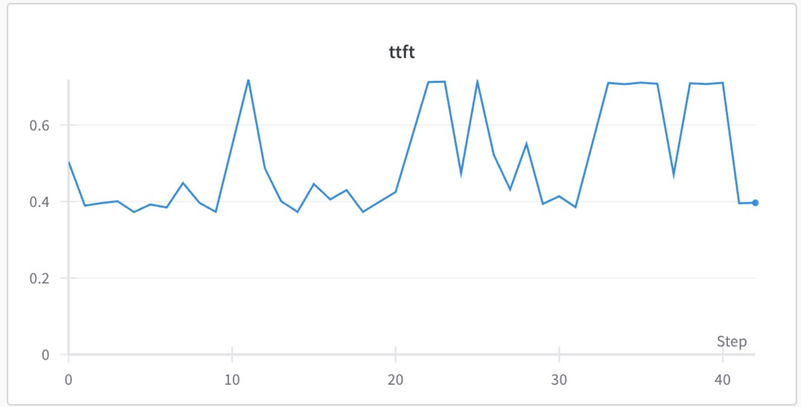
Step 43

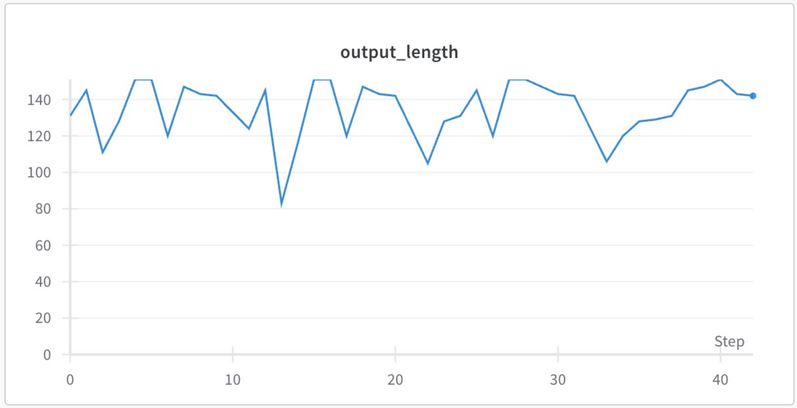
Charts 7

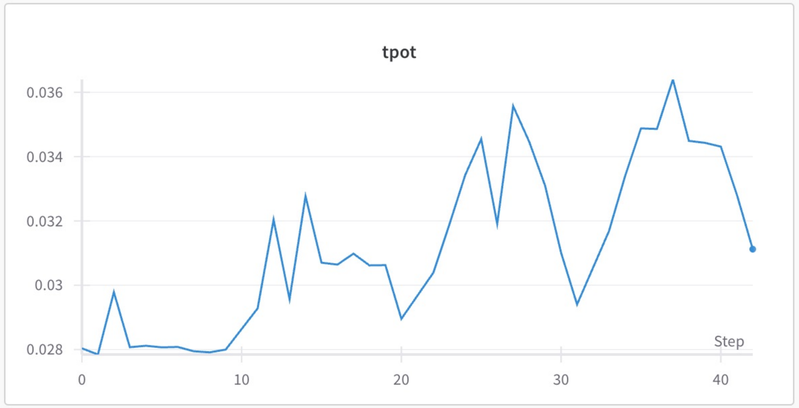
Add panel

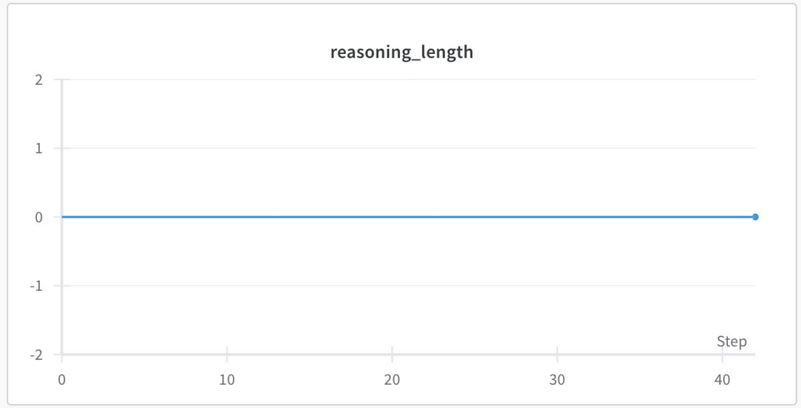








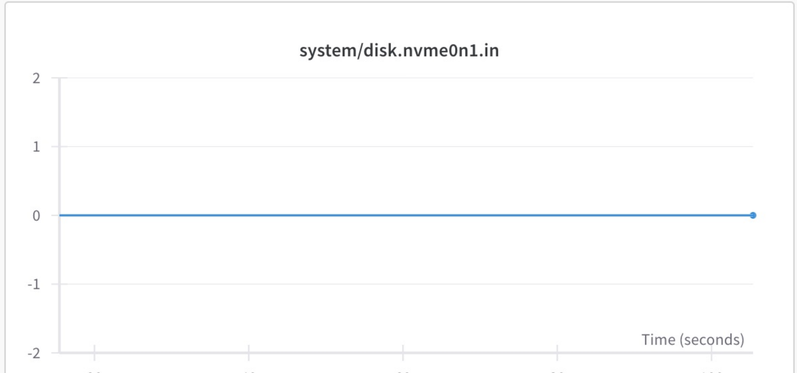


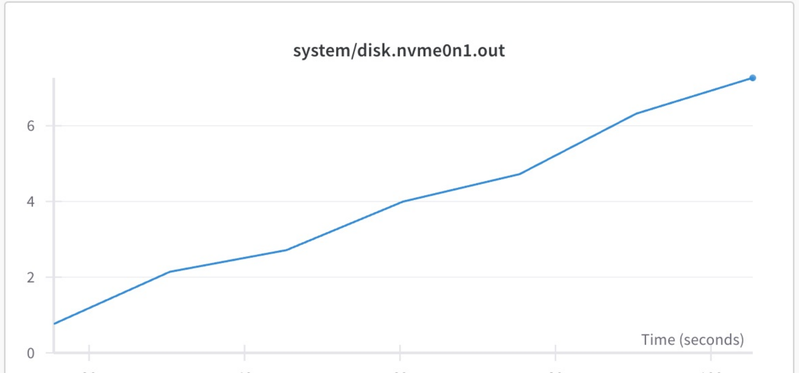


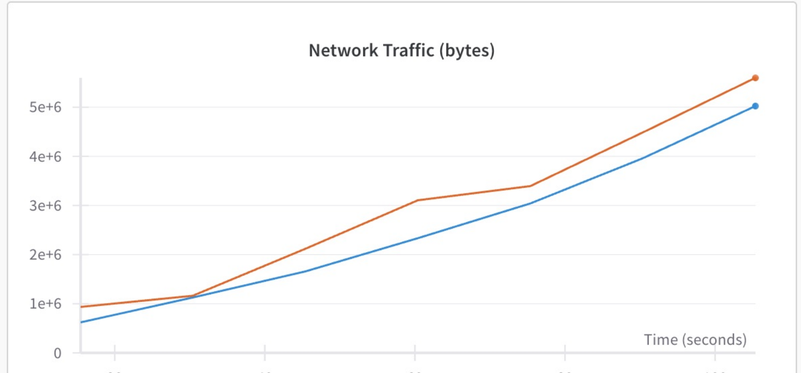
1-6 of 7

System 11

Add panel







Overview
Workspace
Runs
Jobs
Automat.
Sweeps
Reports
Artifacts

Chenyonghua's workspace

Personal workspace

Autosaved just now

Runs (1)

Search runs

Name (1 visualized)

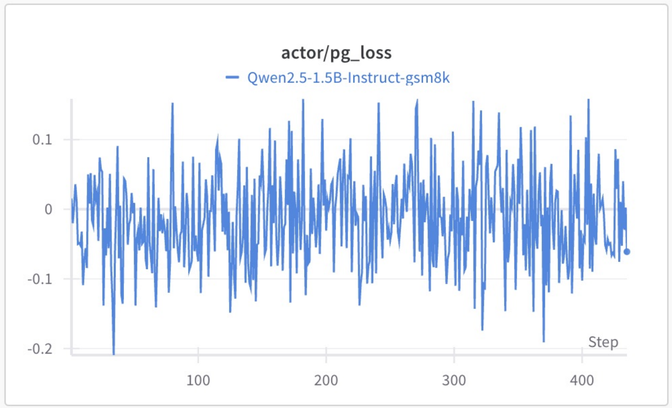
Qwen2.5-1.5...struct-gsm8k

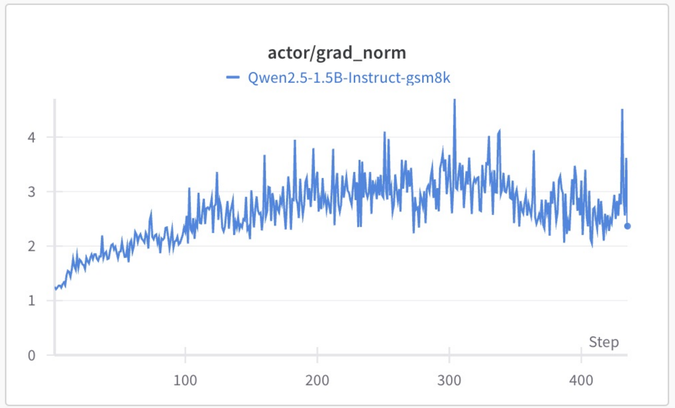
Search panels with regex

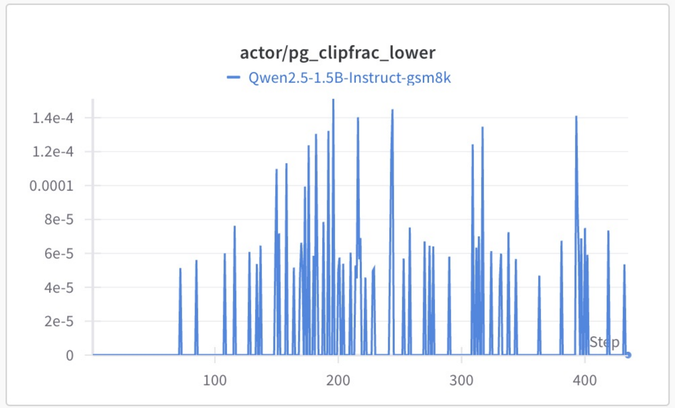
Create report

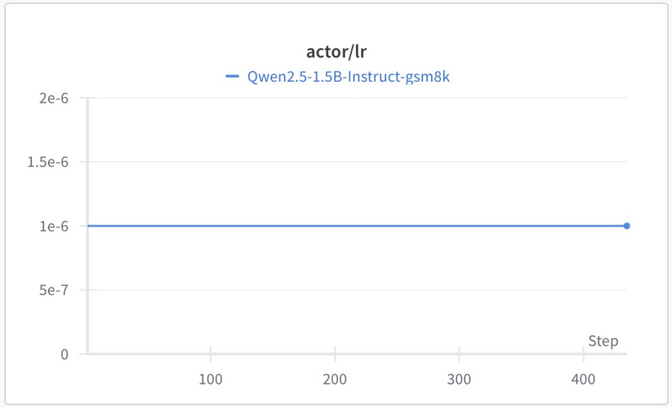
actor 7

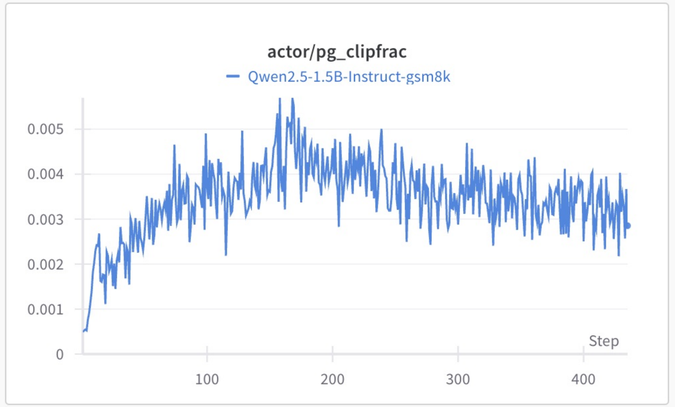
Add panel

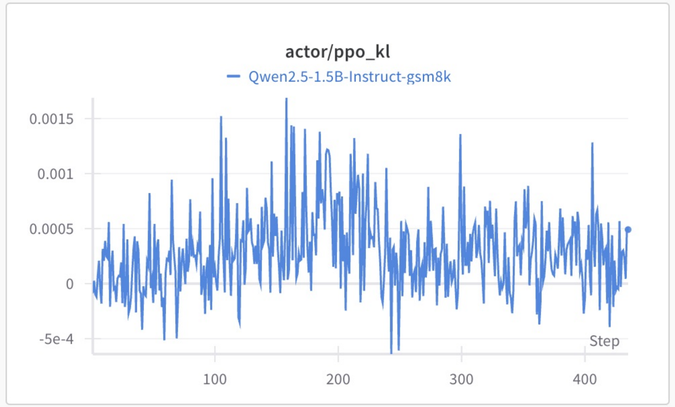








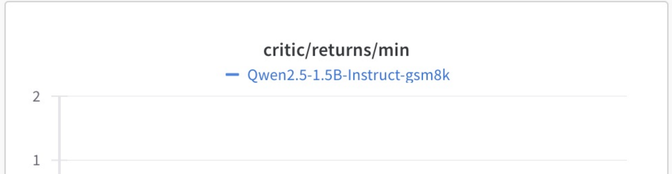


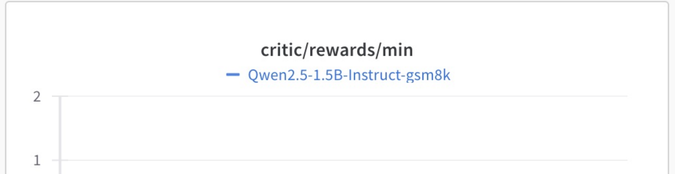


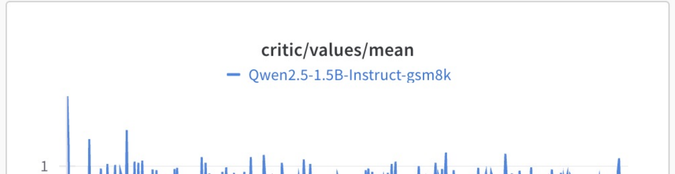
1-6 of 7

critic 21

Add panel







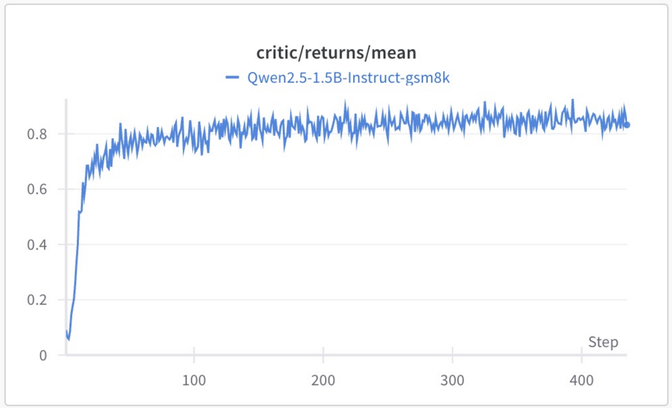
1-1 of 1

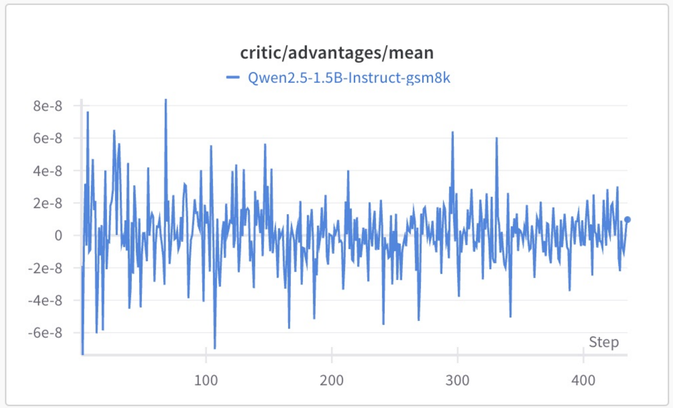
Runs (1)

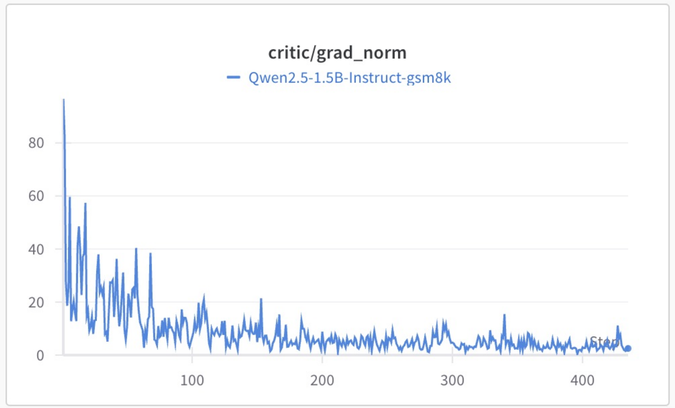
Search runs .*

Name (1 visualized)

Qwen2.5-1.5...struct-gsm8k



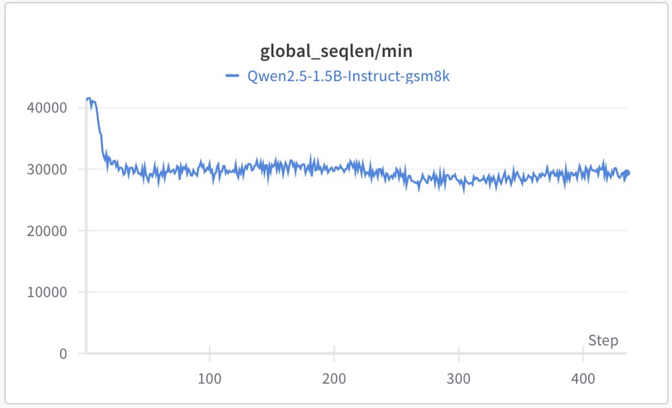


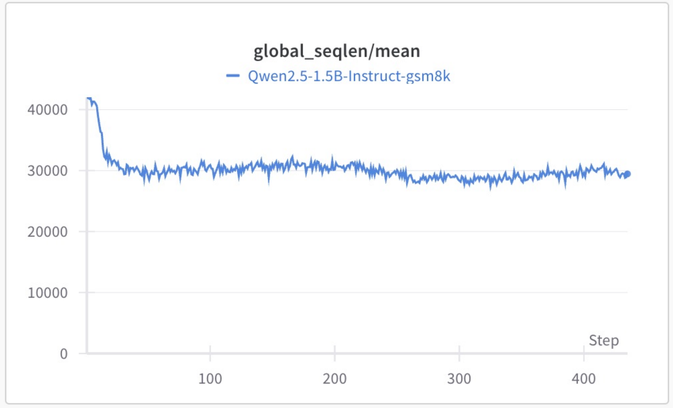


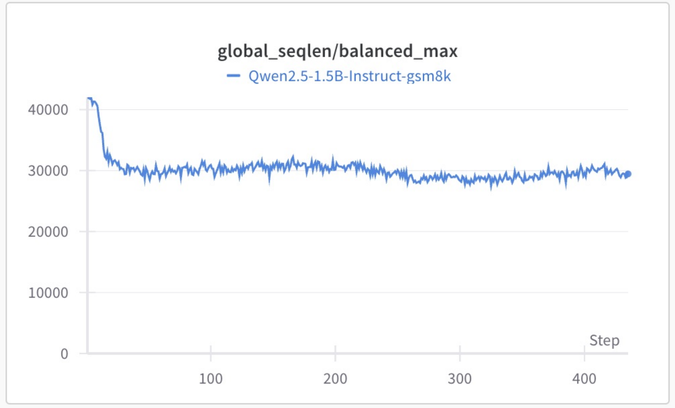
1-6 of 21

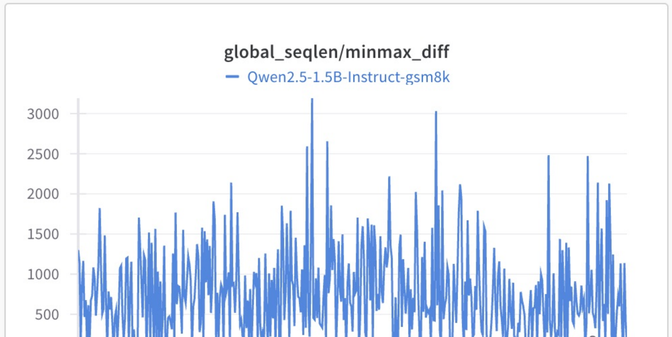
global_seqlen 6

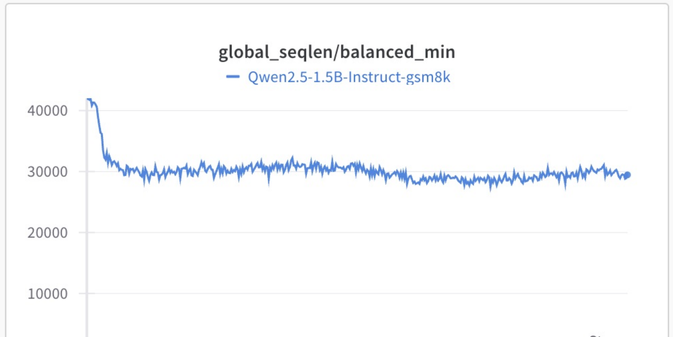
Add panel

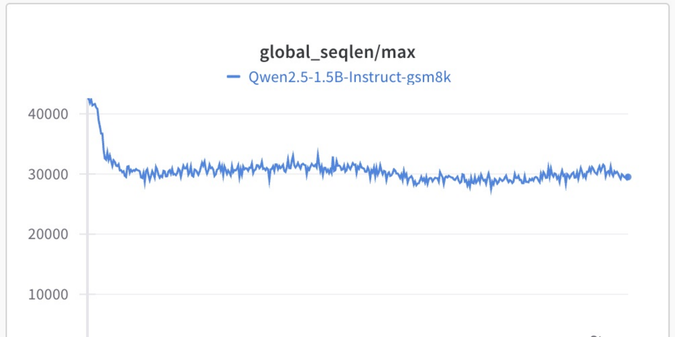












1-1 of 1

Runs (1)

Search runs .*

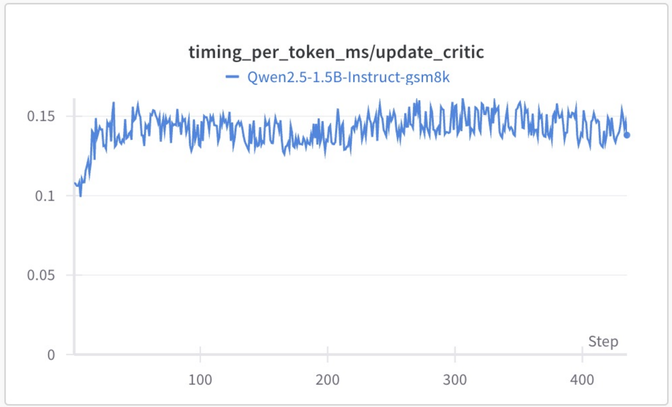
Name (1 visualized)

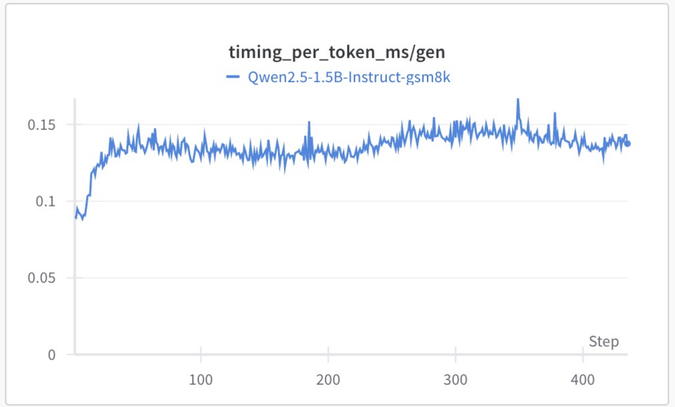
Qwen2.5-1.5...struct-gsm8k

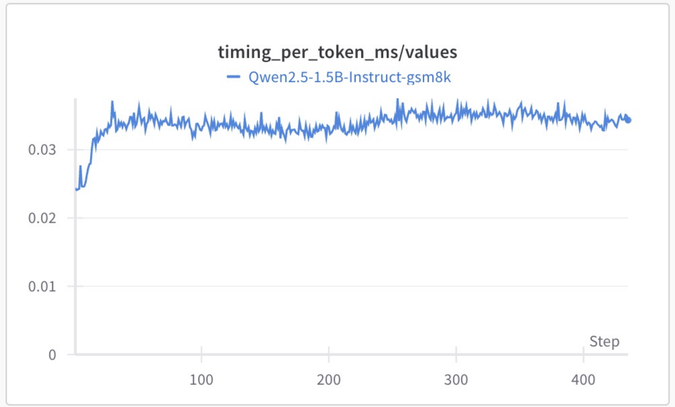
1-1 of 1

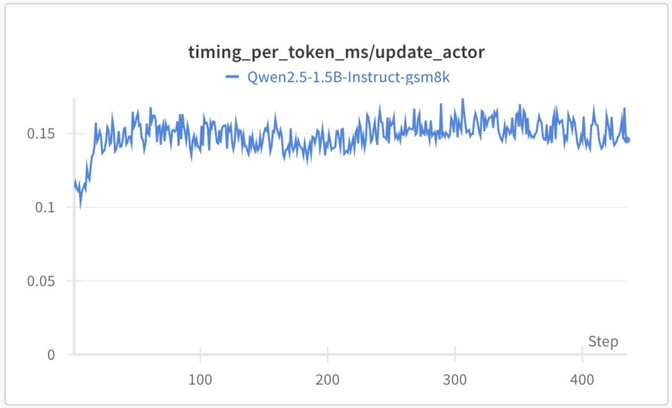
timing_per_token_ms 5

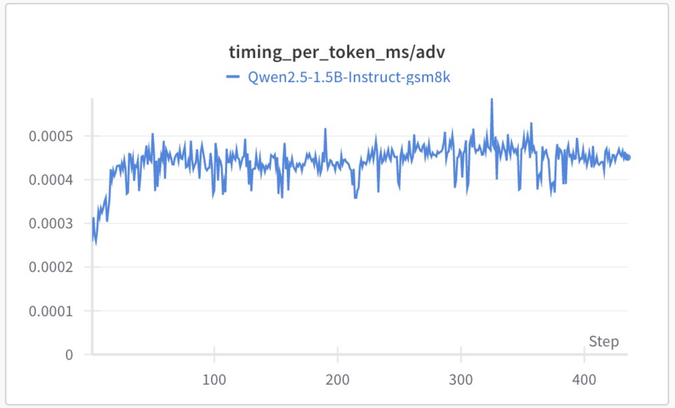
Add panel





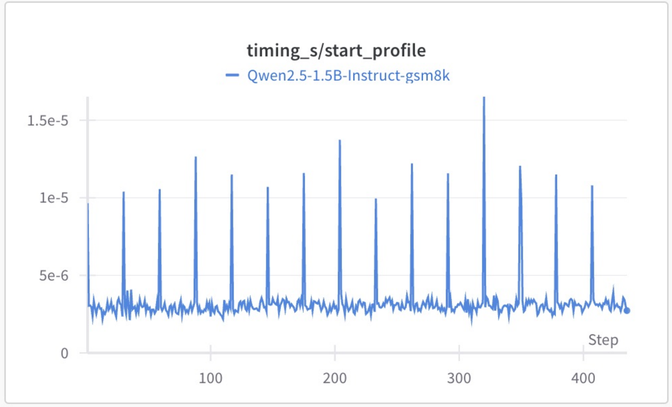


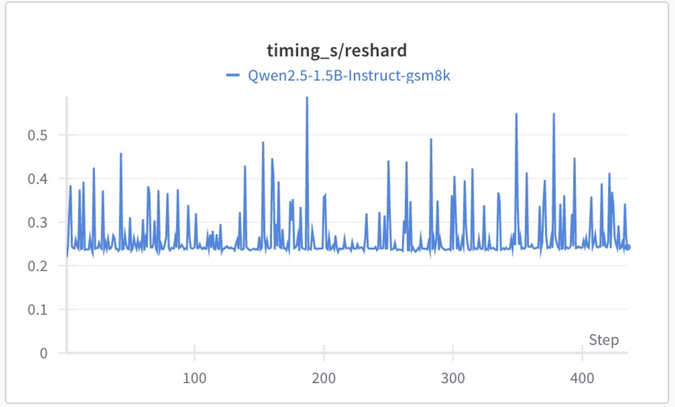


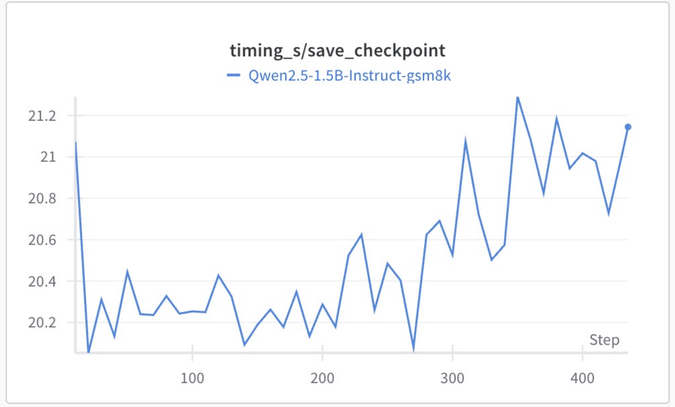


timing_s 14

Add panel





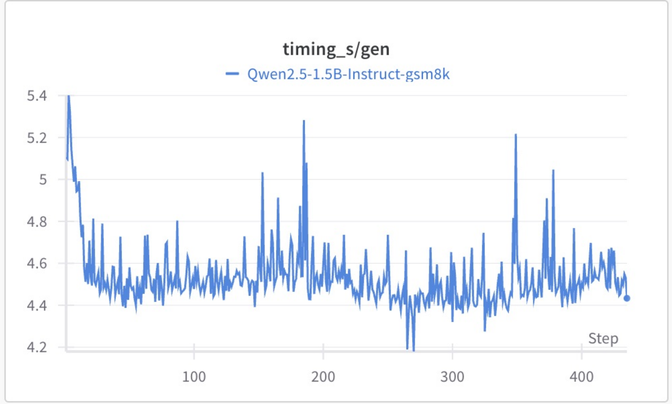


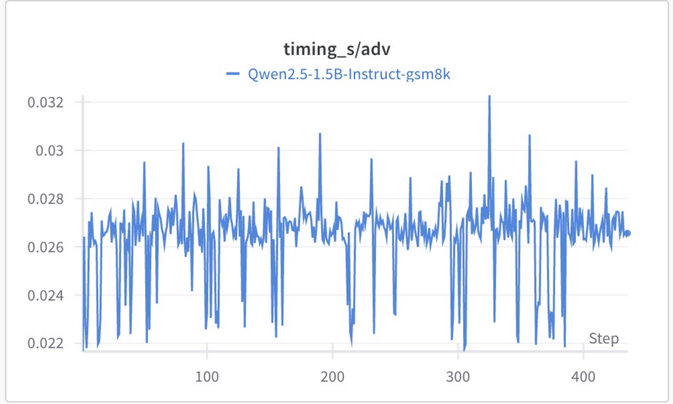
Runs (1)

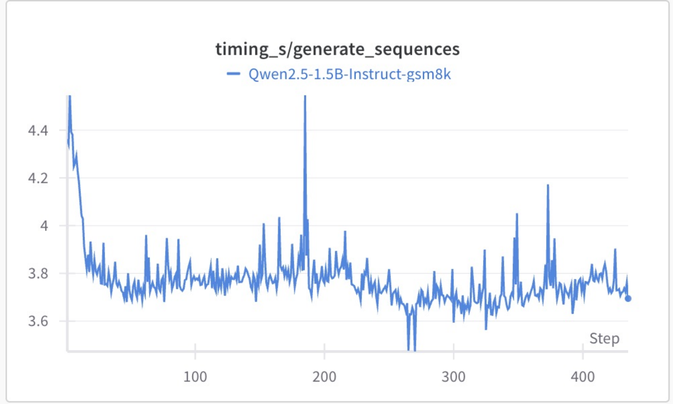
Search runs .*

Name (1 visualized)

Qwen2.5-1.5...struct-gsm8k



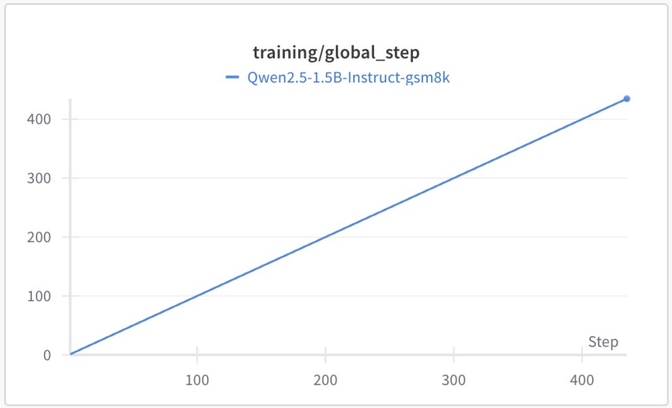


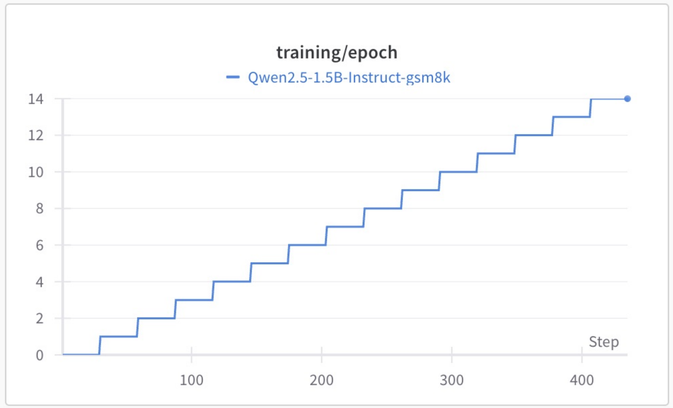


1-6 of 14

training 2

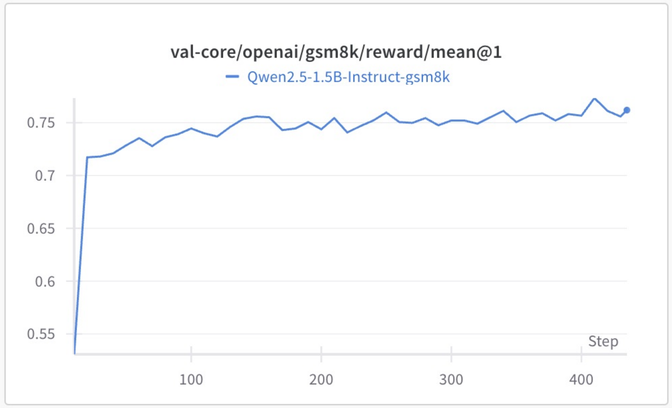
Add panel





val-core 1

Add panel



1-1 of 1