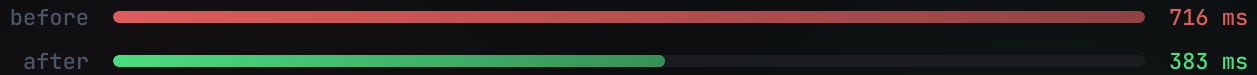


Cold Start vs. Warmed GPU

glasstrace · same model · same workload · device: cuda



● BEFORE 716.27 ms

● AFTER 383.48 ms

```
glasstrace report
modules profiled: 169
total events: 3380
total measured time: 716.27 ms
device: cuda
```

Module	Type	Calls	Total ms	Per-call ms	% of total
model.layers.0.self_attn.q_proj	Linear	20	104.82	5.24	14.6
lm_head	Linear	20	77.73	3.89	10.9
model.layers.0.mlp.gate_proj	Linear	20	14.13	0.71	2
model.layers.15.mlp.gate_proj	Linear	20	8.58	0.43	1.2
model.layers.0.self_attn.o_proj	Linear	20	8.35	0.42	1.2
model.layers.17.mlp.down_proj	Linear	20	5.45	0.27	0.8
model.layers.15.mlp.down_proj	Linear	20	4.92	0.25	0.7
model.layers.16.self_attn.o_proj	Linear	20	4.92	0.25	0.7
model.layers.20.self_attn.v_proj	Linear	20	4.83	0.24	0.7
model.layers.16.self_attn.k_proj	Linear	20	4.78	0.24	0.7
model.layers.0.mlp.down_proj	Linear	20	4.68	0.23	0.7
model.layers.15.self_attn.o_proj	Linear	20	4.65	0.23	0.6
model.layers.23.mlp.down_proj	Linear	20	4.39	0.22	0.6
model.layers.9.mlp.down_proj	Linear	20	4.36	0.22	0.6
model.layers.16.mlp.down_proj	Linear	20	4.33	0.22	0.6
model.layers.21.mlp.down_proj	Linear	20	4.32	0.22	0.6
model.layers.8.mlp.down_proj	Linear	20	4.31	0.22	0.6
model.layers.13.mlp.down_proj	Linear	20	4.25	0.21	0.6
model.layers.2.mlp.down_proj	Linear	20	4.24	0.21	0.6
model.layers.20.mlp.down_proj	Linear	20	4.22	0.21	0.6

```
glasstrace report
modules profiled: 169
total events: 3380
total measured time: 383.48 ms
device: cuda
```

Module	Type	Calls	Total ms	Per-call ms	% of total
lm_head	Linear	20	39.29	1.96	10.2
model.layers.5.mlp.down_proj	Linear	20	3.57	0.18	0.9
model.layers.10.mlp.down_proj	Linear	20	3.54	0.18	0.9
model.layers.0.mlp.down_proj	Linear	20	3.53	0.18	0.9
model.layers.3.mlp.down_proj	Linear	20	3.51	0.18	0.9
model.layers.4.mlp.down_proj	Linear	20	3.51	0.18	0.9
model.layers.8.mlp.down_proj	Linear	20	3.51	0.18	0.9
model.layers.6.mlp.down_proj	Linear	20	3.5	0.17	0.9
model.layers.14.mlp.down_proj	Linear	20	3.5	0.17	0.9
model.layers.11.mlp.down_proj	Linear	20	3.49	0.17	0.9
model.layers.7.mlp.down_proj	Linear	20	3.47	0.17	0.9
model.layers.17.mlp.down_proj	Linear	20	3.47	0.17	0.9
model.layers.1.mlp.down_proj	Linear	20	3.46	0.17	0.9
model.layers.9.mlp.down_proj	Linear	20	3.46	0.17	0.9
model.layers.2.mlp.down_proj	Linear	20	3.45	0.17	0.9
model.layers.13.mlp.down_proj	Linear	20	3.44	0.17	0.9
model.layers.21.mlp.down_proj	Linear	20	3.43	0.17	0.9
model.layers.23.mlp.down_proj	Linear	20	3.43	0.17	0.9
model.layers.18.mlp.down_proj	Linear	20	3.42	0.17	0.9
model.layers.19.mlp.down_proj	Linear	20	3.41	0.17	0.9

TOP HOTSPOT
q_proj · 104.82 ms

LM_HEAD
77.73 ms · 10.9%

TOP HOTSPOT
lm_head · 39.29 ms

SHARE OF TOTAL
10.2% vs 14.6%

● -46.5% total measured time after warm-up