Optimal (Maximum) Batch sizes for Inference on Glue Dataset

Model: Llama-2-7b (NousResearch/Llama-2-7b-hf)

On GPU:

ID	Glue Task	Train size	Optimal batch size	Time for inference on 40 samples **(seconds)**	Estimate of time for inference on all samples **(minutes)**
1	rte	2.49k	4	9.35	9.7
2	cola	8.55k	4	9.38	33
3	mnli	393k	4	9.45	26 **(hrs)**
4	mrpc	3.67k	4	9.38	15
5	qnli	105k	4	9.48	7 **(hrs)**
6	qqp	364k	4	9.58	25 **(hrs)**
7	sst-2	67.3k	4	9.41	4.4 **(hrs)**
8	stsb	5.75k	4	9.46	23
9	wnli	635	4	9.45	2.5

On CPU:

ID	Glue Task	Train size	Optimal batch size	Time for inference on 40 samples **(minutes)**	Estimate of time for inference on all samples **(hours)**
1	rte	2.49k	4	12.0	12.67
2	cola	8.55k	4	9.44	33.63
3	mnli	393k	4	12.40	2030.5
4	mrpc	3.67k	4	12.08	18.47
5	qnli	105k	4	13.10	573.13
6	qqp	364k	4	12.10	1835.17
7	sst-2	67.3k	4	9.75	273
8	stsb	5.75k	4	12.2	29.23
9	wnli	635	4	12.1	3.2