Online Result Summary

## Model: text_reg_batch

GPU(s):

Total Available GPU Memory: 0 GB

Constraint targets: None

In 53 measurements across 9 configurations, **text_reg_batch_config_4** is **50%** better than the default configuration at maximizing throughput, under the given constraints, on GPU(s) .

- **text_reg_batch_config_4**: 0 GPU instance with a max batch size of 16 on platform onnxruntime

Curves corresponding to the 3 best model configuration(s) out of a total of 9 are shown in the plots.
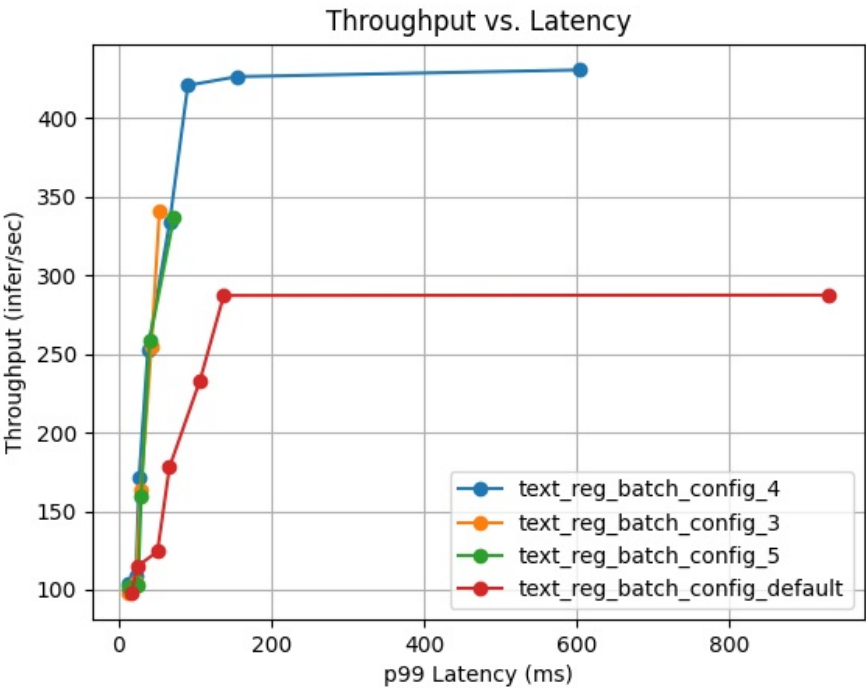
Throughput vs. Latency curves for 3 best configurations.
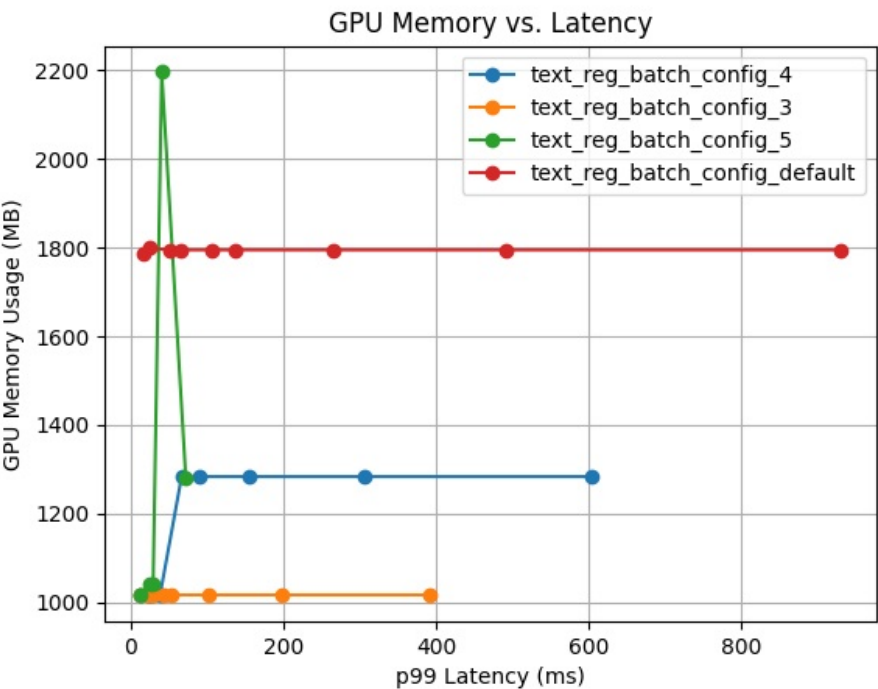
GPU Memory vs. Latency curves for 3 best configurations.

The following table summarizes each configuration at the measurement that optimizes the desired metrics under the given constraints.

Model Config Name	Max Batch Size	Dynamic Batching	Total Instance Count	p99 Latency (ms)	Throughput (infer/sec)	Max GPU Memory Usage (MB)	Average GPU Utilization (%)
text_reg_batch_config_4	16	Enabled	0:GPU	603.053	430.757	1283	94.5
text_reg_batch_config_3	8	Enabled	0:GPU	53.717	341.062	1015	93.5
text_reg_batch_config_5	32	Enabled	0:GPU	72.569	336.716	1279	87.7
text_reg_batch_config_default	8	Enabled	2:GPU	930.668	287.462	1795	74.0