Online Result Summary

Model: florence2_model

GPU(s):

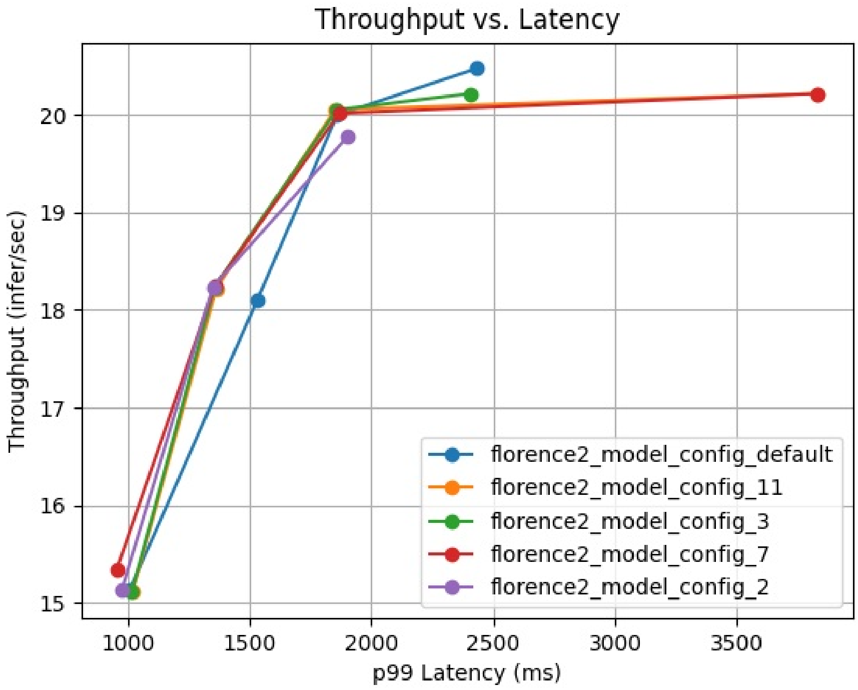
Total Available GPU Memory: 0 GB

Constraint targets: None

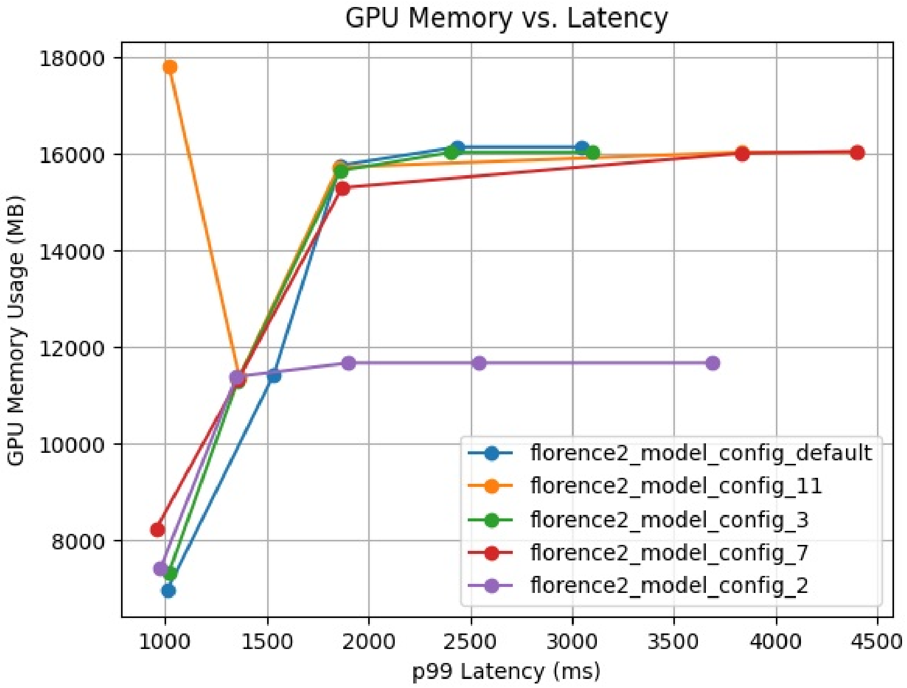
In 65 measurements across 13 configurations, **florence2_model_config_default** provides no gain over the default configuration, under the given constraints, on GPU(s) .

- **florence2_model_config_default**: 0 GPU instance with a max batch size of 36 on platform python

Curves corresponding to the 5 best model configuration(s) out of a total of 13 are shown in the plots.



Throughput vs. Latency curves for 5 best configurations.



GPU Memory vs. Latency curves for 5 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 (%)
florence2_model_config_default	36	Enabled	0:GPU	2431.937	20.4761	16144	75.4
florence2_model_config_11	36	Enabled	0:GPU	3831.547	20.2189	16037	78.5
florence2_model_config_3	36	Enabled	0:GPU	2407.582	20.2184	16035	71.5
florence2_model_config_7	36	Enabled	0:GPU	3831.468	20.218	16016	71.9
florence2_model_config_2	24	Enabled	0:GPU	1901.765	19.7727	11677	71.7