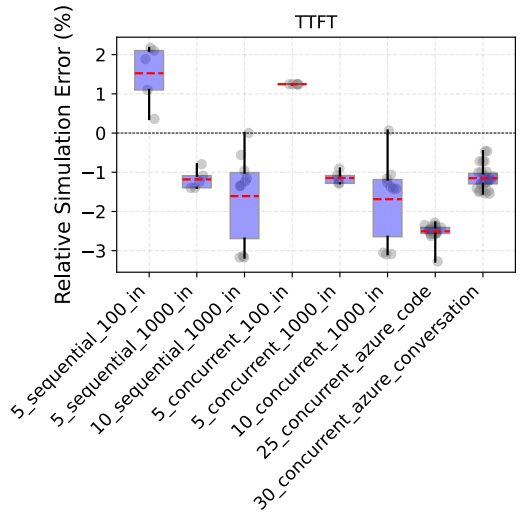

Figure 18: Simulation accuracy for Time-To-First-Token latencies of Llama-3.3 70B in vLLM.

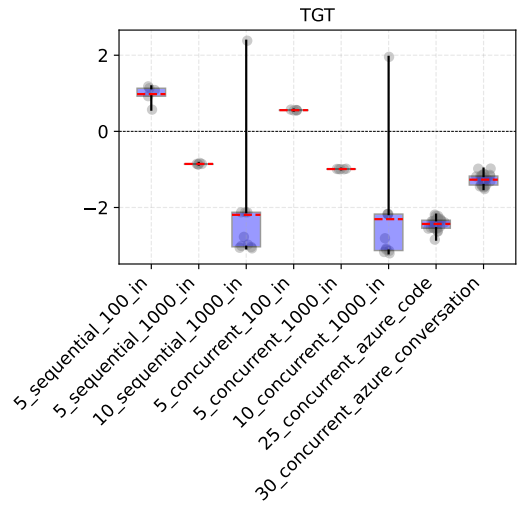

Figure 17: Simulation accuracy for Total-Generation-Time latencies of Llama-3.3 70B in vLLM.

GPUs	Simulated p95 Latency	Measured p95 Latency	Error
10	4.38s	4.37s	0.2%
12	3.36s	3.32s	1.2%
14	3.07s	3.09s	-0.9%
16	2.85s	2.86s	-0.4%

Table 1: Simulated p95 latency for the HellaSwag workload on different numbers of nodes. In the evaluation, each node contains two GPUs (i.e., it involves 5, 6, 7, 8 nodes).

A Simulator

Ken uses a simulator to evaluate cascades and gear plans during the offline phase. Prior work has shown that the repetitive and deterministic nature of performing the linear algebra inside a machine learning model can accurately be simulated [31, 48]. Ken follows the approach of previous work and uses a continuous-time, discrete-event simulator. We use the same simulator in to evaluate gear plans in the offline phase and to run theexperiments in Section 5.5 and Section 11.

Table 1 shows the simulation accuracy of our simulator on top of the ExLlama engine running Llama-2 models as per the workload described in section 5.1. The simulator is very accurate in this scenario since each query involves a single forward pass with a large model, leaving little room for noise or error accumulations.

Figures 18 and 17 show the simulation accuracy on top of vLLM. While the simulator is sufficiently accurate for our purposes, it cannot be as accurate as when simulating ExLlama as vLLM involves more complicated logic where requests involve several autoregressive forward passes and several requests are processed inside the same forward pass.

Big, big thanks to Sarah Wang who built and evaluated the vLLM simulator.