# A TERMINOLOGY

*Table 1.* Summary of metrics used throughout the paper.

| Metric | Acronym | Definition |
|---|---|---|
| First Token Latency | FTL | Latency to run prefill and generate the first token. |
| Token-to-Token Latency | TTL | Latency to generate each new token in decoding. |
| Tokens per Second per User | TPS | The rate of token generation per user ($1/\text{TTL}$). |
| Service Level Agreement | SLA | Agreed upon P50 TTL and FTL for a service. |
| Batch (a.k.a., concurrency) | – | The number of requests a system can serve per model replica. |
| Context Throughput (per GPU) | – | The throughput per context (prefill) GPU expressed in requests/second/GPU. This accounts for only the context (prefill) GPUs deployed. |
| Decode Throughput (per GPU) | – | The throughput per decode (generation) GPU expressed in tokens/second/GPU. This accounts for only the decode (generation) GPUs deployed. |
| Overall Throughput (per GPU) | – | The total throughput of the system expressed in tokens/second/GPU. This accounts for all GPUs deployed (prefill and decode). |
| Utilization | – | For a hardware resource, the percentage of execution time that the resource would be busy assuming 100% efficiency. |

# B USING 50TH PERCENTILE (P50) STATISTICS AS PROXY FOR PERFORMANCE ANALYSIS

Figure 16 presents the CDF of input and output sequence lengths (ISL and OSL) from a real-world deployed workload. Absolute values have been obfuscated for privacy; however, the shape of the distribution is representative.

Figure 17 shows the resulting Pareto frontiers when this traffic distribution is simulated directly versus when it is approximated using a simplified approach: the closest power-of-two values to the P50 of ISL and OSL. Notably, the approximated frontier closely matches the original, indicating that using P50 ISL and OSL as an approximation provides a reasonable overview of the trends.

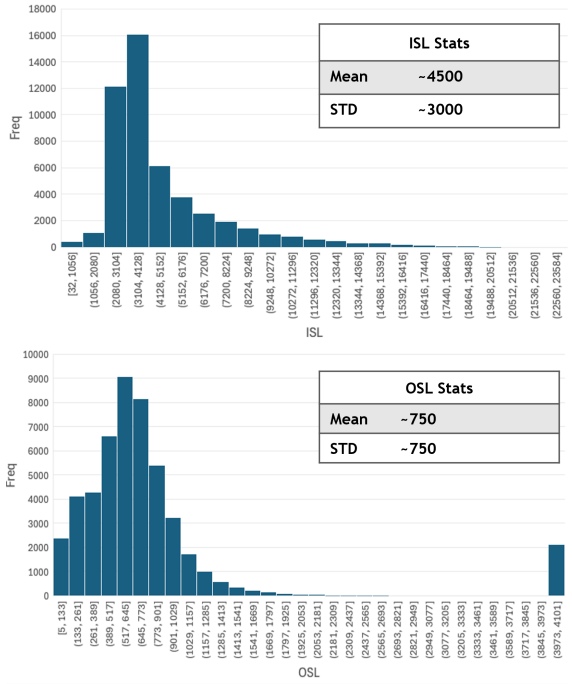


*Figure 16.* Example distribution of ISL and OSL in dynamic traffic.

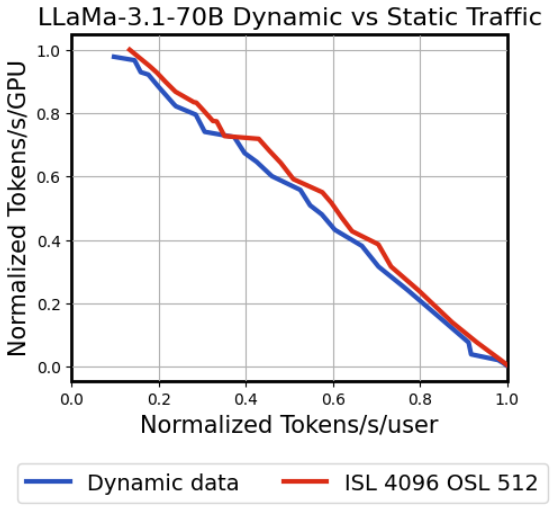


*Figure 17.* Comparison of Pareto frontiers using dynamic traffic simulation versus P50 approximation.