CHAPTER 17 / LATENCY

First token is not finished.

Separate waiting, prompt processing and output generation.

Separate waiting, prompt processing and output generation.



Illustrative timing model: 100 ms network plus queue, prefill and (N−1)/rate. No real service measurements or guaranteed speeds.

TTFT

Network, queueing and prompt processing contribute to the first visible token.

Completion

After the first token, additional output steps extend total response time.

Serving

Latency, throughput, batching and cache behavior need workload-specific measurement.

↗ **Network, queueing and prompt processing contribute to the first visible token.**

PUT IT TO WORK

With the same compute timeline, streaming shows the first token earlier; it does not shorten the generation interval.

CLOSE THE PAGE. EXPLAIN IT.

Would you reduce the queue or the output length to improve this user's main wait?