

Inference Saturation Curve

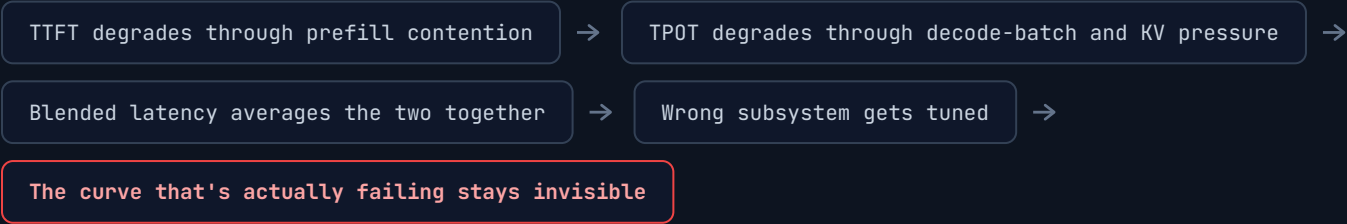
DEFINITION

The nonlinear relationship between inference load and latency/throughput as a serving stack approaches saturation.

FRAMEWORK FLOW

01 The Condition TTFT and TPOT each degrade through different mechanisms as concurrency rises — prefill contention on one side, decode-batch and KV-cache pressure on the other.	02 The Boundary Both curves hold near-linear behavior up to a concurrency threshold, then bend — the point where each subsystem's headroom runs out.	03 Failure State A single blended latency number hides which curve is actually driving degradation, so the wrong subsystem gets tuned.	04 Consequence Capacity and batching decisions get made against an average neither curve individually supports, and the fix targets the wrong layer.
---	---	---	---

Failure State: One Number Hides Two Curves



THE DIAGNOSTIC SIGNAL

SEPARATE TTFT / TPOT MODELING Prefill and decode saturate on independent curves — collapsing them into one number obscures which is binding.	SHAPE OVER POINT The value is in the curve's shape and knee location, not a single precise concurrency figure.
FIRST SATURATION DRIVER The named subsystem — KV ceiling, decode saturation, prefill amplification, or queue amplification — that bends the curve first.	DIRECTIONAL, NOT PREDICTIVE Every model constant is exposed; the curve is a directional signal under stated assumptions, not a forecast.

ARCHITECTURAL RELATIONSHIPS

#95 Interaction Collapse Point RELATED MODERATE Co-resident in the AI Inference Saturation Analyzer — the Interaction Collapse Point sits on the Inference Saturation Curve.	#97 Token Queue Amplification RELATED MODERATE Co-resident in the Saturation Analyzer — queue amplification is the mechanism shaping the saturation curve.
#98 Throughput Illusion RELATED MODERATE Co-resident in the Saturation Analyzer as paired curve/illusion metrics.	

ANCHOR /ai-inference-saturation-analyzer/

FRAMEWORK HOME

TTFT and TPOT modeled as separate degradation curves, with the collapse knee marked and every assumption exposed.

"A single blended latency number hides which subsystem is actually binding first — separating the curves is what lets the failure be named."