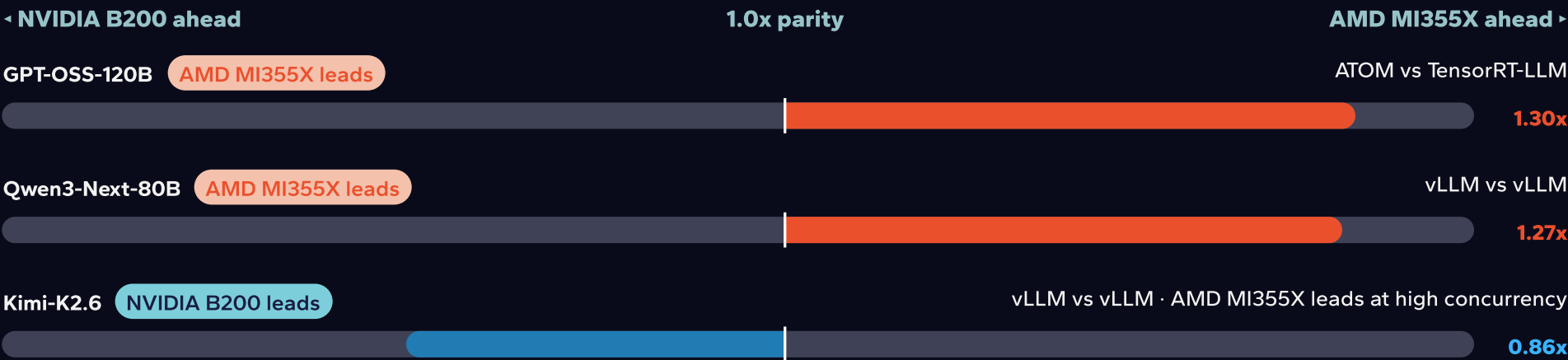


AMD Instinct MI355X vs NVIDIA HGX B200

A three-level inference infrastructure evaluation — raw performance, cost, and real business impact, tested across three production LLMs.

Performance

AMD MI355X leads on 2 of 3 models tested. NVIDIA B200 leads only on Kimi-K2.6 at low-to-moderate concurrency — AMD MI355X pulls back ahead at high concurrency.



Economics

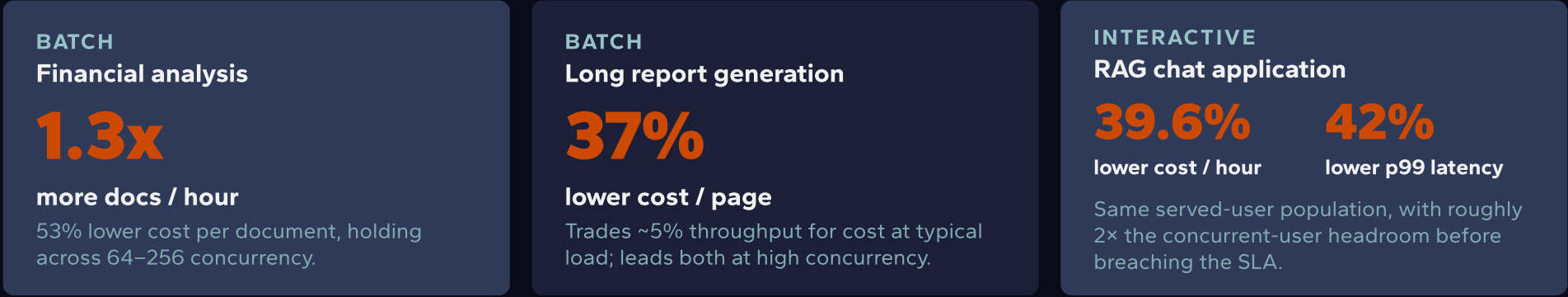
AMD Instinct MI355X achieves consistent cost advantages over NVIDIA HGX B200, based on blended on-demand pricing across cloud providers.¹ MI355X testing was hosted via TensorWave, an AMD-focused cloud provider.



1. Cloud GPU pricing collected July 10, 2026 across four providers; rates fluctuate with market demand.

Business Impact

AMD MI355X's performance and cost advantages translate into real business impact — demonstrated here by mapping token throughput to three representative enterprise workloads.



Across every dimension tested — **performance, cost, and business outcomes** — AMD MI355X delivers the stronger result, even where NVIDIA is faster in raw throughput.