Scaling Validation Report v3

Key Findings:

- 8 replicas held 56.14 req/s at 400 users (0 failures) after ramp/timeout tuning; +13.
- The earlier debug run logged 25.95% failures; rerun now completes 10,111 calls wit
- CPU snapshots show 95-154% utilisation per inference container (avg ~130%), indi
- 200-user / 8-replica sweep remains warm-up limited (20.44 req/s, 0 fails) and is trea

Artifacts:

- reports/scaling_validation_v3/scaling_metrics.csv (and .json)
- reports/scaling_validation_v3/scaling_metadata.csv (and .json)
- reports/scaling_validation_v3/rps_vs_replicas.[png|pdf] (plus median/p95/failure cha
- locust/results/multi_replica_v1/replicas8_u400/ (clean rerun + CPU stats)
- locust/results/multi_replica_v1/replicas8_u400_obsolete/ (superseded artifacts)

Next Steps:

- Optional: capture the 800-user / 8-replica lane with -r 30 and refresh visuals.