

Text Inference Capacity Review

Committee-facing summary of the controlled simultaneous conversation benchmark executed on May 7, 2026. Scope covered the local Mac mini text stack and the DGX Spark text stack under mixed-model load.

| | |
|------------------|---|
| Prepared for | Technical Review Committee |
| Prepared by | Codex benchmark automation in the PerkNation workspace |
| Report date | May 07, 2026 |
| Source artifacts | Benchmark results, per-request logs, and environmental samples captured under a controlled staircase load |

Executive Summary

- Both units remained reachable and stable throughout the benchmark. No request failures, no host loss, and no thermal shutdown conditions were observed.
- The Mac mini safely supported the full controlled test band through **4 simultaneous mixed-model conversations**. At that point it delivered **0.265 rps** with **26.08s** P95 latency and no thermal warnings.
- The DGX Spark stayed stable through **8 simultaneous mixed-model conversations**, but useful throughput flattened near **0.135 rps** by the 2-conversation band. Higher concurrency primarily increased queueing delay rather than completed work.
- Spark P95 latency rose from **21.07s** at 1 concurrent conversation to **127.35s** at 8, while GPU temperature stayed controlled at **67.00C** peak.
- The Mac mini result is conservative because Low Power Mode was enabled during the run, which likely understates its actual ceiling under a performance-oriented configuration.

Mac mini operating recommendation. Validated from 1 to 4 concurrent mixed-model conversations with zero failures. Operate at up to **4 simultaneous live conversations** today. Treat any higher ceiling as likely but not yet formally validated. Committee read: healthy and unsaturated inside the tested band; measured result is conservative because Low Power Mode was enabled.

DGX Spark operating recommendation. Validated from 1 to 8 concurrent mixed-model conversations with zero failures. Use **2 simultaneous conversations** for premium responsiveness, **4** if slower replies are acceptable, and **6 to 8** only when the application exposes queueing expectations. Committee read: stable under load, but effectively compute-bound after the 2-conversation band.

Test Design And Safeguards

The benchmark used direct Ollama-compatible `/v1/chat/completions` calls against each unit with a short multi-turn prompt payload. Each unit was tested separately to avoid cross-contamination. Concurrency was increased in steps rather than as a single spike.

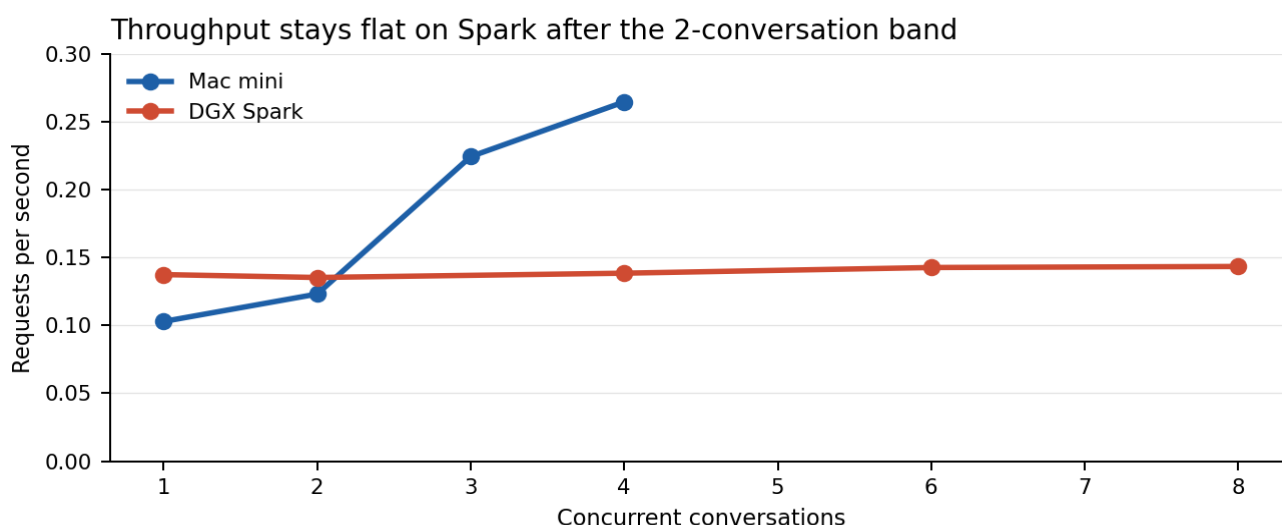
- Mac mini planned staircase: 1, 2, 3, 4 concurrent conversations.
- DGX Spark planned staircase: 1, 2, 4, 6, 8 concurrent conversations.
- Total requests per step were sized as $\max(\text{concurrency} \times 4, \text{model_count} \times 3)$ so each band exercised all models repeatedly.
- Safety stops were active. The Mac run would halt on thermal or performance warnings. The Spark run would halt on any request errors or if GPU temperature reached 72C.
- The Mac test was intentionally capped at 4 to preserve access and avoid risking workstation usability during business hours.

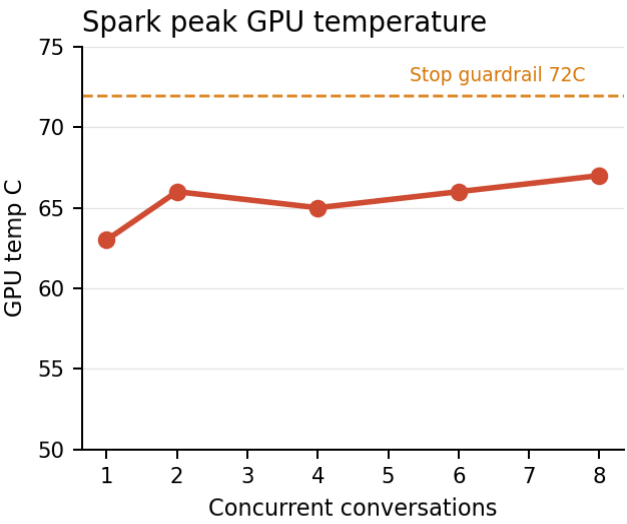
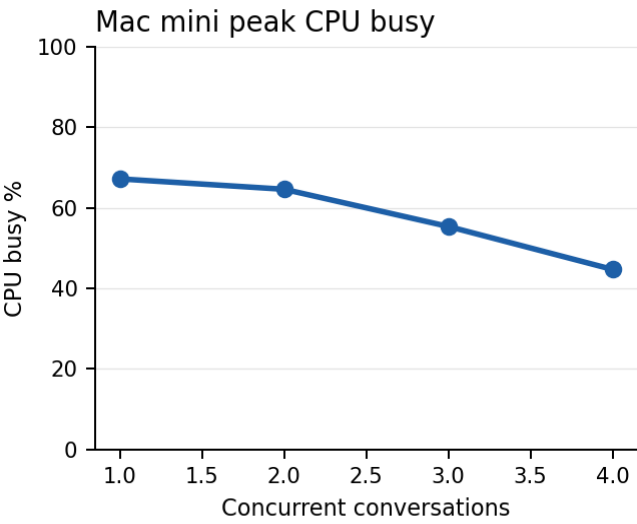
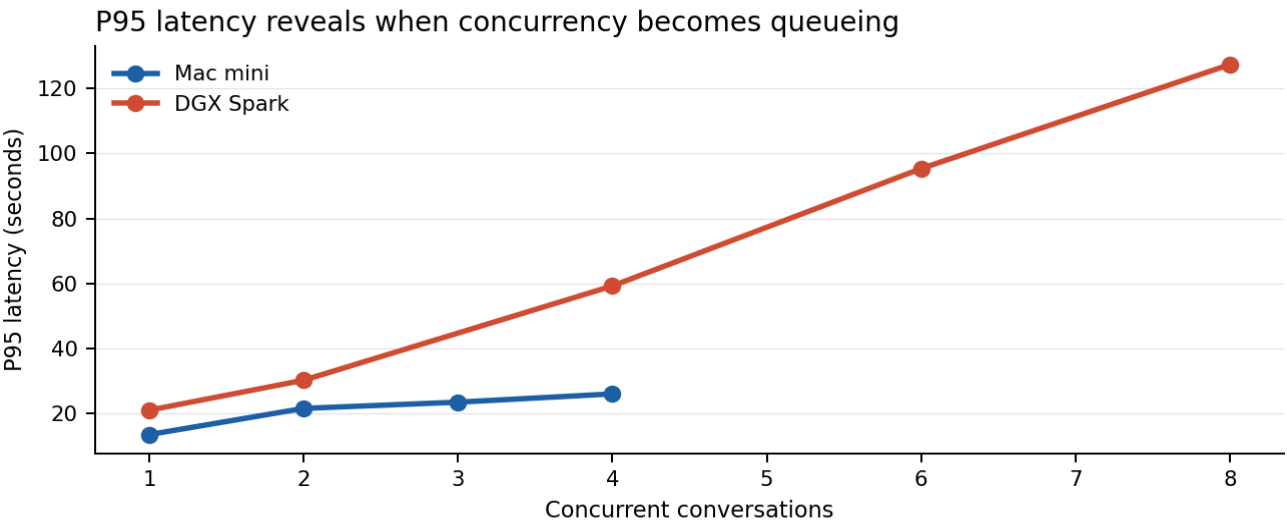
Mac mini profile. Apple M4 Pro with 12 CPU cores (8 performance + 4 efficiency), 48 GB RAM, macOS 26.3. Models under test: qwen2.5:7b, qwen3:30b, nemotron-3-nano:30b. Telemetry captured: CPU busy %, load average, memory free %, and macOS thermal-warning state. Important caveat: Low Power Mode was enabled during the run.

DGX Spark profile. NVIDIA DGX Spark running Ubuntu 24.04.4 with an ARM 20-core CPU and about 122 GiB of system memory. Models under test: llava:latest, nemotron-mini:latest, nemotron:latest, qwen3:8b. Telemetry captured: GPU temperature, GPU utilization, GPU power draw, load average, and memory used in MB.

Interpretive note: “maxed out” in this report means the point where additional simultaneous conversations stop creating meaningful throughput gains and instead translate into waiting time. It does not only mean a host crash or thermal shutdown.

Performance Curves





Findings By Platform

Mac mini. The workstation remained healthy through the full tested envelope. Throughput improved at each higher step, culminating at 0.265 rps at 4 concurrent conversations. The absence of thermal and performance warnings indicates the box was not pressured into a protective state.

Operational reading. Four concurrent mixed-model text conversations are supportable today. Because the test ended at 4 by design, this is a validated floor for capacity, not the absolute ceiling.

| Concurrency | Success | Throughput rps | Avg latency s | P95 s | Peak CPU % | Load1 | Thermal warn |
|-------------|---------|----------------|---------------|-------|------------|-------|--------------|
| 1 | 9/9 | 0.103 | 9.72 | 13.57 | 67.18 | 7.07 | No |
| 2 | 9/9 | 0.123 | 14.90 | 21.62 | 64.60 | 6.68 | No |
| 3 | 12/12 | 0.224 | 11.22 | 23.54 | 55.35 | 7.28 | No |
| 4 | 16/16 | 0.265 | 13.39 | 26.08 | 44.67 | 5.91 | No |

DGX Spark. The server remained stable all the way to 8 concurrent conversations. However, throughput stayed almost flat between 1 and 8 concurrent sessions, moving only from 0.137 to 0.143 rps, while P95 latency rose from 21.07s to 127.35s. This is the signature of saturation-by-queueing rather than failure.

Operational reading. Spark can host more simultaneous conversations than the Mac mini, but only the first 2 to 4 feel interactive. Beyond that, users wait longer without receiving proportionally more completed work.

| Concurrency | Success | Throughput rps | Avg latency s | P95 s | Peak GPU C | Peak GPU % | Peak GPU W |
|-------------|---------|----------------|---------------|--------|------------|------------|------------|
| 1 | 12/12 | 0.137 | 7.28 | 21.07 | 63.00 | 96.00 | 49.61 |
| 2 | 12/12 | 0.135 | 13.33 | 30.28 | 66.00 | 96.00 | 46.61 |
| 4 | 16/16 | 0.138 | 22.31 | 59.27 | 65.00 | 96.00 | 48.01 |
| 6 | 24/24 | 0.143 | 29.83 | 95.35 | 66.00 | 96.00 | 47.52 |
| 8 | 32/32 | 0.143 | 38.15 | 127.35 | 67.00 | 96.00 | 47.16 |

Committee Conclusions

1. The Mac mini is currently the safer choice for a small, interactive text surface where predictable response times matter and concurrency stays at or below four active conversations.
2. The DGX Spark is the stronger shared compute node when the application can tolerate queueing, backgrounding, or explicit “working” indicators. It does not fail early, but it stops getting materially faster after the 2-conversation band.
3. If the goal is premium user experience, the current scheduling policy should treat Spark as a 2-conversation premium lane and a 4-conversation soft ceiling. Higher concurrency should be admitted only when the UI makes the waiting state explicit.
4. The Mac mini result should not be treated as a fixed hard cap. The run was intentionally conservative, and Low Power Mode was enabled, so a follow-up off-hours sweep at 5 and 6 concurrent conversations is likely to find additional headroom.

Recommended Actions

- Keep the Mac mini in the current controlled duty band until a follow-up test is scheduled. If a higher ceiling is needed, rerun with Low Power Mode disabled and extend the staircase to 6.
- On Spark, use request queueing and visible progress indicators rather than allowing users to infer responsiveness from concurrency alone.
- If Spark is expected to host premium interactive chat, route the slowest large model class away from the shared queue or apply per-model admission controls.
- Preserve the benchmark harness and rerun after any model inventory change, host power-profile change, or driver/runtime update.

Limitations

- This was a mixed-model test, not a per-model isolation study. Individual model ceilings may differ materially from the aggregate result.
- The Mac mini did not expose a numeric silicon temperature without elevated privileges. The report therefore uses macOS thermal-warning state rather than die temperature.
- Spark was measured with direct GPU thermal telemetry, but the report does not separate GPU time by model family. A per-model pass would sharpen scheduling policy.
- Request content was intentionally short and operational. Longer prompts or larger context windows would reduce the measured ceilings.

Supporting files: raw logs, environmental samples, and the benchmark harness are preserved alongside this PDF in the benchmark artifact directory for committee review and reproducibility.