Spark Open-Weight Model Bakeoff

Committee benchmark comparing Nemotron Super, GLM-4.5-Air Q6_K, and Kimi Dev 72B Q4_0 on the live DGX Spark lane.

Scope: serial per-model benchmark across short ops chat, structured JSON, and long-context synthesis, followed by a controlled dual-Ollama cold-load probe.

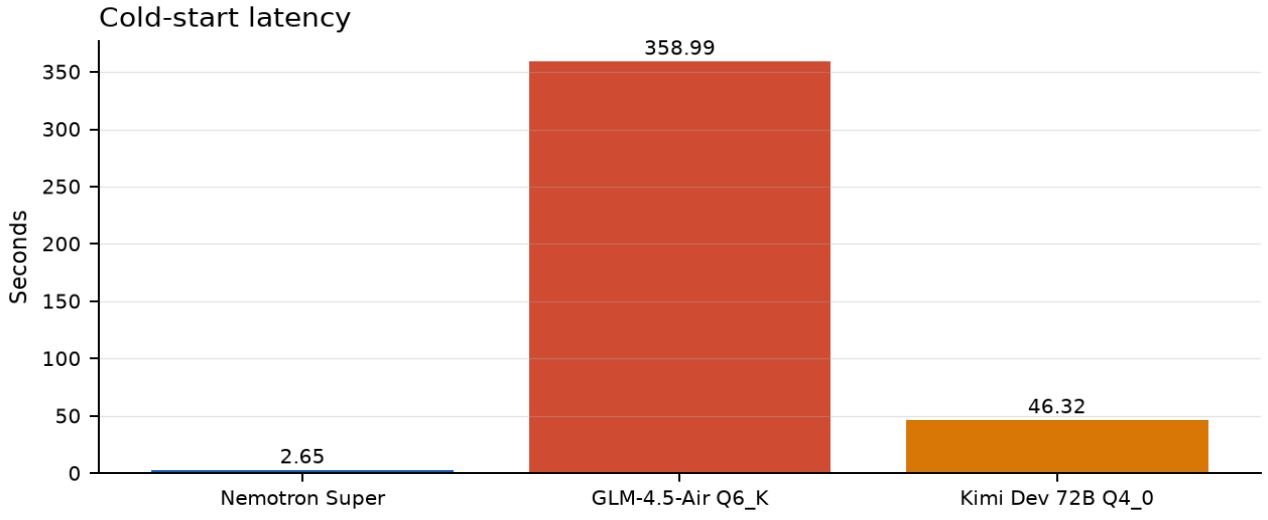
Executive read. Nemotron is the public default. GLM remains the stronger comparison lane for operators who need a second open model on Spark. Kimi stayed runnable but is materially slower and belongs in a specialty slot with lower concurrency expectations.

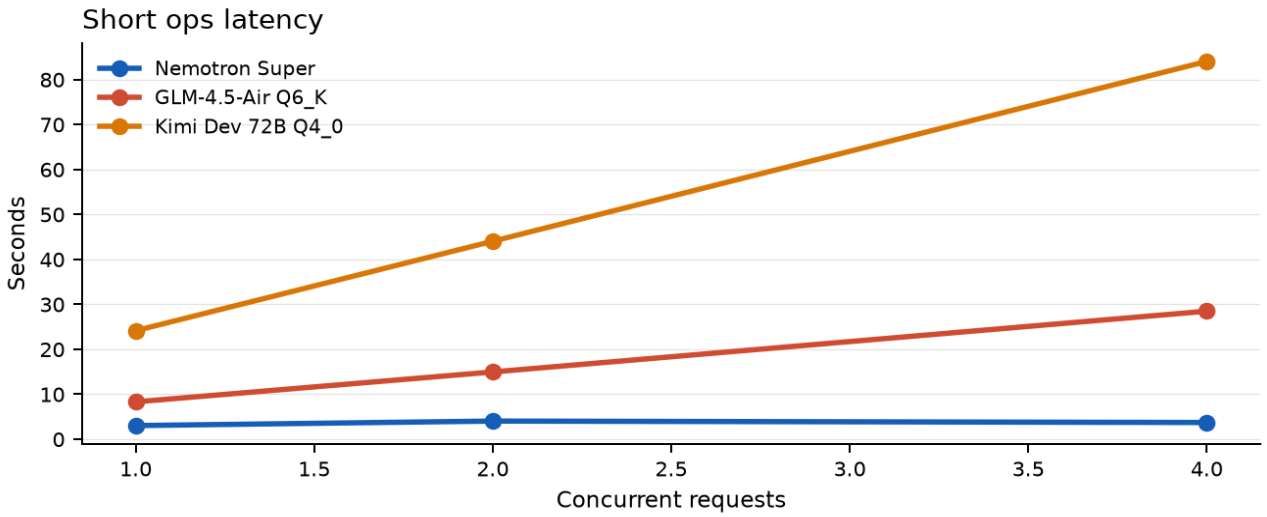
Headline findings

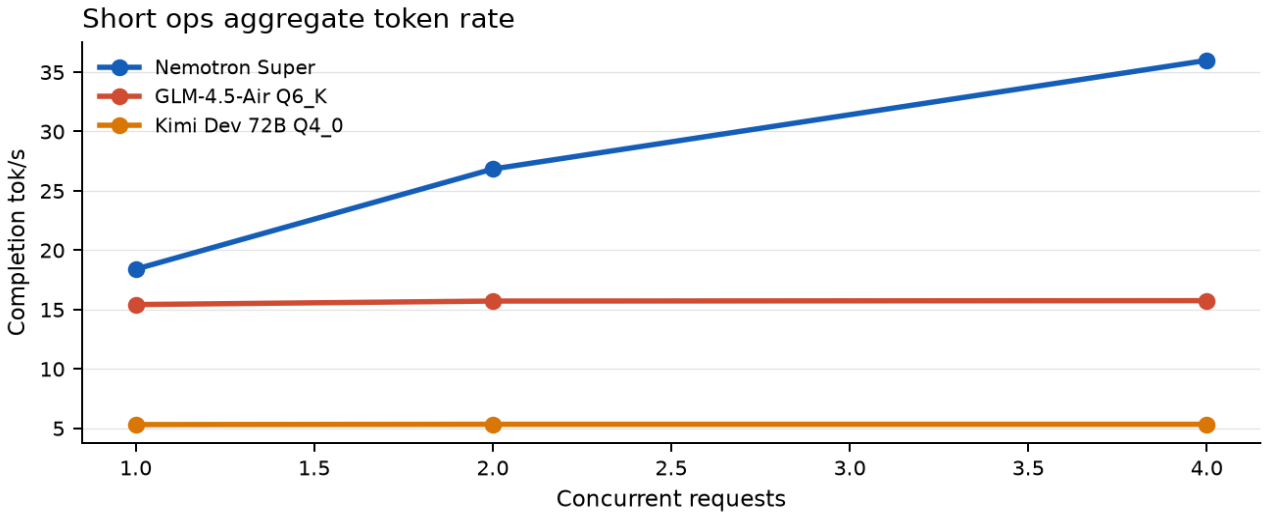
Fastest cold start: **Nemotron Super** at 2.65 seconds. Slowest cold start: **GLM-4.5-Air Q6_K** at 358.99 seconds. The dual-Ollama probe bottomed out at 27481 MB available RAM with n/a MB of swap used.

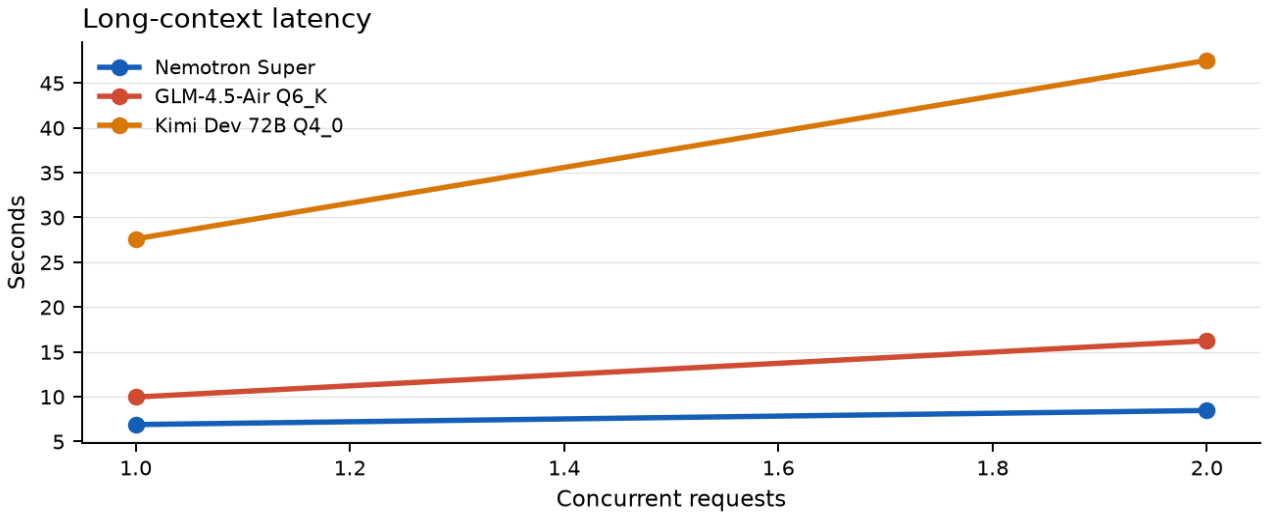
Live-stack limitation. The co-resident GLM attempt on 2026-07-27T19:04:26Z failed before inference with **CUDA error: out of memory while TRT services were resident**. This is why the GLM and Kimi ladders were rerun in an isolated Ollama window after the TRT services were stopped.

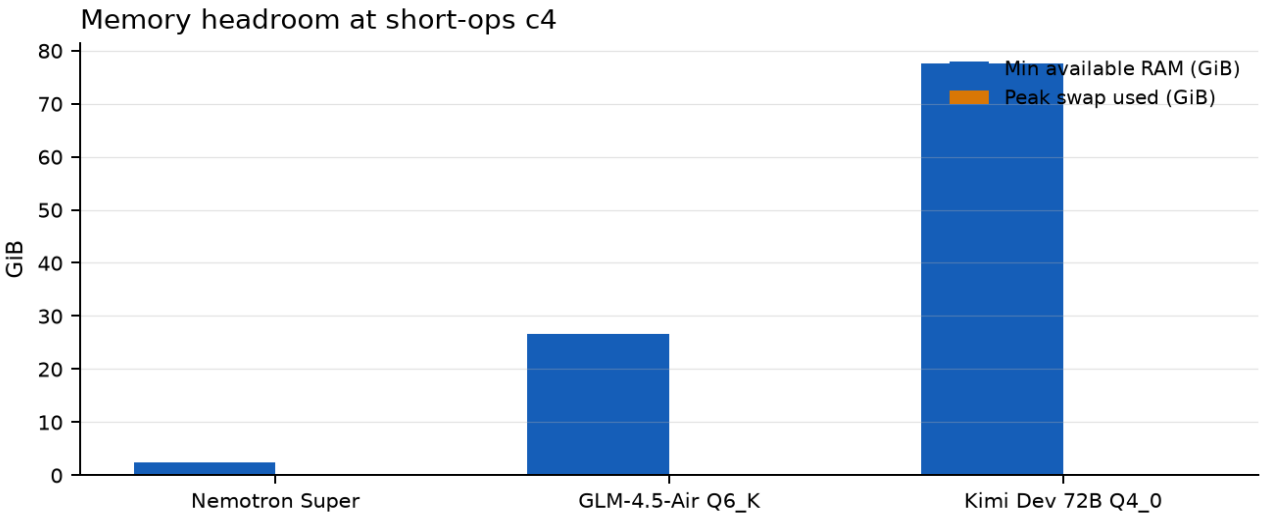
The unsupported Kimi Linear image was removed from Spark before the benchmark, reclaiming 52 GB and removing a misleading catalog entry that could not be executed by the local Ollama runtime.




Short ops latency
Nemotron Super
GLM-4.5-Air Q6_K
Kimi Dev 72B Q4_0
Seconds
0
10
20
30
40
50
60
70
80
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Concurrent requests


Short ops aggregate token rate
Nemotron Super
GLM-4.5-Air Q6_K
Kimi Dev 72B Q4_0
Completion tok/s
5
10
15
20
25
30
35
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Concurrent requests


Long-context latency
Nemotron Super
GLM-4.5-Air Q6_K
Kimi Dev 72B Q4_0
Seconds
5
10
15
20
25
30
35
40
45
1.0
1.2
1.4
1.6
1.8
2.0
Concurrent requests


Memory headroom at short-ops c4
GiB
0
10
20
30
40
50
60
70
80
Min available RAM (GiB)
Peak swap used (GiB)
Nemotron Super
GLM-4.5-Air Q6_K
Kimi Dev 72B Q4_0

Operator recommendation

Default public lane: Nemotron Super. It is the fastest model in cold start and steady-state throughput, and it avoids the shared-Ollama contention story entirely because it stays on the dedicated TensorRT-LLM stack.

Secondary comparison lane: GLM-4.5-Air Q6_K. It is materially slower to cold-load than Nemotron but still meaningfully faster than Kimi once resident. It is the best Spark-side comparison lane when a second open-weight text model is needed.

Specialty lane: Kimi Dev 72B Q4_0. It stayed functional, but the slower latency curve and the dual-Ollama memory pressure make it a specialty path rather than a first-choice public default.

Model	c1 avg s	c2 avg s	c4 avg s	c4 tok/s	c4 free MB	c4 swap
Nemotron Super	3.02	4.03	3.70	35.98	2380	n/a
GLM-4.5-Air Q6_K	8.30	14.95	28.47	15.74	27189	n/a
Kimi Dev 72B Q4_0	24.13	44.02	84.04	5.33	79393	n/a

Dual-Ollama cold-load probe: 2/2 requests succeeded. Average latency was 191.61 seconds and wall time was 354.27 seconds.

Artifacts included with this report: raw JSON, CSV request logs, sampled host vitals, the benchmark runner used for the sweep, and the rendered committee PDF.