

www.phoronix-test-suite.com

deepsparse 7950x

AMD Ryzen 9 7950X 16-Core testing with a ASUS ROG CROSSHAIR X670E HERO (0805 BIOS) and AMD Radeon RX 6800/6800 XT / 6900 on Ubuntu 22.10 via the Phoronix Test Suite.

Automated Executive Summary

c had the most wins, coming in first place for 33% of the tests.

Based on the geometric mean of all complete results, the fastest (a) was 1.002x the speed of the slowest (c). bb was 0.999x the speed of a, d was 1x the speed of bb, c was 1x the speed of d.

The results with the greatest spread from best to worst included:

Neural Magic DeepSparse (Model: NLP Text Classification, DistilBERT mnli - Scenario: Asynchronous Multi-Stream) at 1.016x
Neural Magic DeepSparse (Model: NLP Text Classification, DistilBERT mnli - Scenario: Asynchronous Multi-Stream) at 1.016x
Neural Magic DeepSparse (Model: CV Detection, YOLOv5s COCO - Scenario: Asynchronous Multi-Stream) at 1.014x
Neural Magic DeepSparse (Model: CV Detection, YOLOv5s COCO - Scenario: Asynchronous Multi-Stream) at 1.014x
Neural Magic DeepSparse (Model: NLP Text Classification, BERT base uncased SST2 - Scenario: Synchronous Single-Stream) at 1.013x
Neural Magic DeepSparse (Model: NLP Text Classification, BERT base uncased SST2 - Scenario: Synchronous

Single-Stream) at 1.013x
Neural Magic DeepSparse (Model: NLP Question Answering, BERT base uncased SQuaD 12layer Pruned90 - Scenario: Synchronous Single-Stream) at 1.011x
Neural Magic DeepSparse (Model: NLP Question Answering, BERT base uncased SQuaD 12layer Pruned90 - Scenario: Synchronous Single-Stream) at 1.011x
Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Synchronous Single-Stream) at 1.009x
Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Synchronous Single-Stream) at 1.009x.

Test Systems:

a

Processor: AMD Ryzen 9 7950X 16-Core @ 5.88GHz (16 Cores / 32 Threads), Motherboard: ASUS ROG CROSSHAIR X670E HERO (0805 BIOS), Chipset: AMD Device 14d8, Memory: 32GB, Disk: 2048GB SOLIDIGM SSDPFKKW020X7 + 257GB Flash Drive, Graphics: AMD Radeon RX 6800/6800 XT / 6900, Audio: AMD Navi 21/23, Monitor: ASUS MG28U, Network: Intel I225-V + Intel Wi-Fi 6 AX210/AX211/AX411

OS: Ubuntu 22.10, Kernel: 5.19.0-29-generic (x86_64), Desktop: GNOME Shell 43.1, Display Server: X Server + Wayland, Vulkan: 1.3.224, Compiler: GCC 12.2.0, File-System: ext4, Screen Resolution: 3840x2160

Kernel Notes: Transparent Huge Pages: madvise
Processor Notes: Scaling Governor: amd-pstate schedutil (Boost: Enabled) - CPU Microcode: 0xa601203
Python Notes: Python 3.10.7
Security Notes: itlb_multihit: Not affected + l1tf: Not affected + mds: Not affected + meltdown: Not affected + mmio_stale_data: Not affected + retbleed: Not affected + spec_store_bypass: Mitigation of SSB disabled via prctl + spectre_v1: Mitigation of usercopy/swapgs barriers and __user pointer sanitization + spectre_v2: Mitigation of Retpolines IBPB: conditional IBRS_FW STIBP: always-on RSB filling PBRSB-eIBRS: Not affected + srbds: Not affected + tsx_async_abort: Not affected

bb

c

d

Processor: AMD Ryzen 9 7950X 16-Core @ 5.88GHz (16 Cores / 32 Threads), Motherboard: ASUS ROG CROSSHAIR X670E HERO (0805 BIOS), Chipset: AMD Device 14d8, Memory: 32GB, Disk: 2048GB SOLIDIGM SSDPFKKW020X7 + 257GB Flash Drive, Graphics: AMD Radeon RX 6800/6800 XT / 6900 (2575/1000MHz), Audio: AMD Navi 21/23, Monitor: ASUS MG28U, Network: Intel I225-V + Intel Wi-Fi 6 AX210/AX211/AX411

OS: Ubuntu 22.10, Kernel: 5.19.0-29-generic (x86_64), Desktop: GNOME Shell 43.1, Display Server: X Server + Wayland, OpenGL: 4.6 Mesa 22.2.1 (LLVM 15.0.2 DRM 3.47), Vulkan: 1.3.224, Compiler: GCC 12.2.0, File-System: ext4, Screen Resolution: 3840x2160

Kernel Notes: Transparent Huge Pages: madvise
Processor Notes: Scaling Governor: amd-pstate schedutil (Boost: Enabled) - CPU Microcode: 0xa601203
Python Notes: Python 3.10.7
Security Notes: itlb_multihit: Not affected + l1tf: Not affected + mds: Not affected + meltdown: Not affected + mmio_stale_data: Not affected + retbleed: Not affected + spec_store_bypass: Mitigation of SSB disabled via prctl + spectre_v1: Mitigation of usercopy/swapgs barriers and __user pointer sanitization + spectre_v2: Mitigation of Retpolines IBPB: conditional IBRS_FW STIBP: always-on RSB filling PBRSB-eIBRS: Not affected + srbds: Not affected + tsx_async_abort: Not affected

deepsparse 7950x

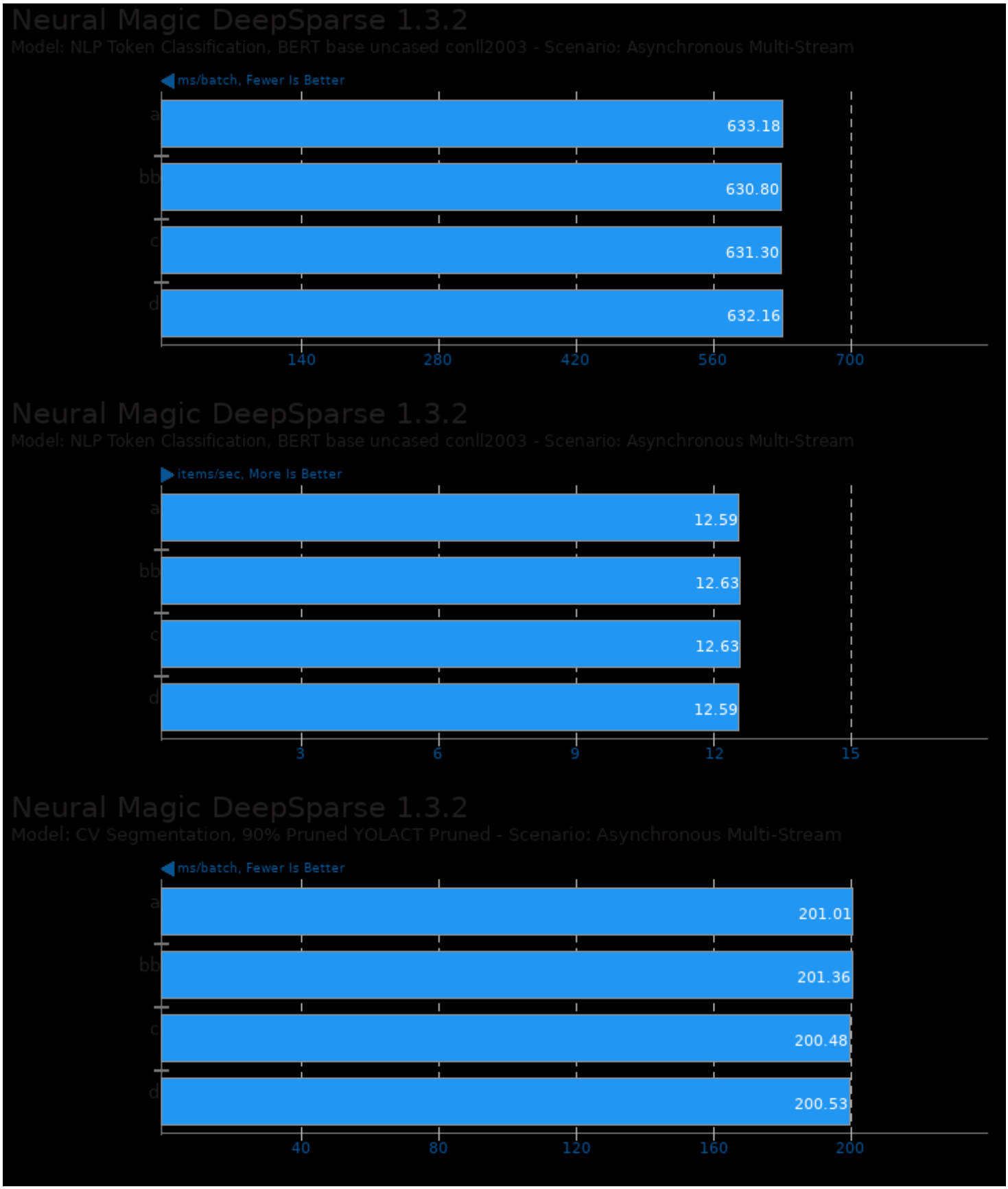
	a	bb	c	d
Neural Magic DeepSparse - N.T.C.B.b.u.c - A.M.S (ms/batch)	633.1822	630.7995	631.3046	632.1566
Normalized	99.62%	100%	99.92%	99.79%
Neural Magic DeepSparse - N.T.C.B.b.u.c - A.M.S (items/sec)	12.5896	12.6314	12.6256	12.593
Normalized	99.67%	100%	99.95%	99.7%
Neural Magic DeepSparse - C.S.9.P.Y.P - A.M.S (ms/batch)	201.0063	201.3574	200.4828	200.5266
Normalized	99.74%	99.57%	100%	99.98%
Neural Magic DeepSparse - C.S.9.P.Y.P - A.M.S (items/sec)	39.7599	39.6409	39.8652	39.8529
Normalized	99.74%	99.44%	100%	99.97%
Neural Magic DeepSparse - N.T.C.B.b.u.S - A.M.S (ms/batch)	83.27	83.6973	83.0764	83.6639
Normalized	99.77%	99.26%	100%	99.3%
Neural Magic DeepSparse - N.T.C.B.b.u.S - A.M.S (items/sec)	96.0573	95.567	96.2813	95.6055
Normalized	99.77%	99.26%	100%	99.3%
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - A.M.S (ms/batch)	627.4365	630.5156	632.9304	630.3581
Normalized	100%	99.51%	99.13%	99.54%
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - A.M.S (items/sec)	12.7031	12.6369	12.5918	12.6421
Normalized	100%	99.48%	99.12%	99.52%
Neural Magic DeepSparse - C.S.9.P.Y.P - S.S.S (ms/batch)	29.4789	29.4922	29.4905	29.5119
Normalized	100%	99.95%	99.96%	99.89%
Neural Magic DeepSparse - C.S.9.P.Y.P - S.S.S (items/sec)	33.9089	33.8938	33.8956	33.8714
Normalized	100%	99.96%	99.96%	99.89%
Neural Magic DeepSparse - N.T.C.B.b.u.S - S.S.S (ms/batch)	16.1956	16.3534	16.1524	16.1441
Normalized	99.68%	98.72%	99.95%	100%
Neural Magic DeepSparse - N.T.C.B.b.u.S - S.S.S (items/sec)	61.7312	61.1354	61.896	61.9274
Normalized	99.68%	98.72%	99.95%	100%
Neural Magic DeepSparse - N.T.C.D.m - S.S.S (ms/batch)	8.5625	8.5043	8.5719	8.5044
Normalized	99.32%	100%	99.21%	100%
Neural Magic DeepSparse - N.T.C.D.m - S.S.S (items/sec)	116.7361	117.5369	116.6075	117.534
Normalized	99.32%	100%	99.21%	100%
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - S.S.S (ms/batch)	85.0258	85.3311	85.5102	85.3242
Normalized	100%	99.64%	99.43%	99.65%
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - S.S.S (items/sec)	11.7605	11.7185	11.694	11.7194
Normalized	100%	99.64%	99.43%	99.65%

deepsparse 7950x

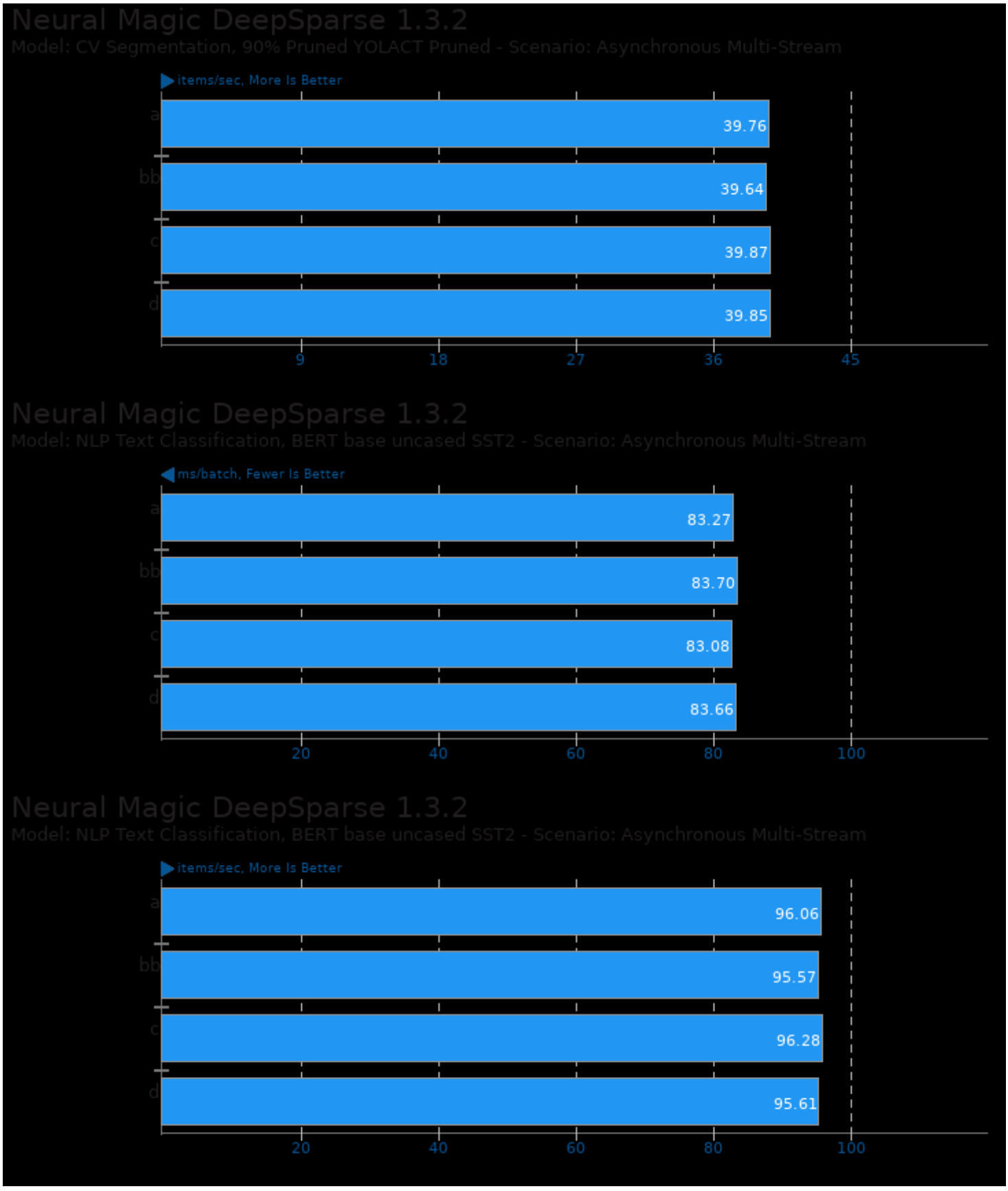
Neural Magic DeepSparse - N.T.C.B.b.u.c - S.S.S (ms/batch)	85.3482	**85.3815**	**85.1808**	85.2218
Normalized	99.8%	99.76%	100%	99.95%
Neural Magic DeepSparse - N.T.C.B.b.u.c - S.S.S (items/sec)	11.7161	**11.7116**	**11.7392**	11.7336
Normalized	99.8%	99.76%	100%	99.95%
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - A.M.S (ms/batch)	78.9798	**78.736**	**79.376**	79.2372
Normalized	99.69%	100%	99.19%	99.37%
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - A.M.S (items/sec)	101.2583	**101.5713**	**100.7526**	100.9291
Normalized	99.69%	100%	99.19%	99.37%
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - A.M.S (ms/batch)	**17.8715**	**17.9286**	17.8979	17.8947
Normalized	100%	99.68%	99.85%	99.87%
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - A.M.S (items/sec)	**447.3079**	**445.8868**	446.6482	446.7292
Normalized	100%	99.68%	99.85%	99.87%
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - S.S.S (ms/batch)	6.1019	6.1208	**6.1393**	**6.0829**
Normalized	99.69%	99.38%	99.08%	100%
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - S.S.S (items/sec)	163.7662	163.2539	**162.7819**	**164.2746**
Normalized	99.69%	99.38%	99.09%	100%
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - S.S.S (ms/batch)	**15.4177**	**15.5803**	15.5237	15.512
Normalized	100%	98.96%	99.32%	99.39%
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - S.S.S (items/sec)	**64.8383**	**64.1578**	64.3933	64.4436
Normalized	100%	98.95%	99.31%	99.39%
Neural Magic DeepSparse - N.T.C.D.m - A.M.S (ms/batch)	**43.2539**	43.3682	43.7142	**43.9364**
Normalized	100%	99.74%	98.95%	98.45%
Neural Magic DeepSparse - N.T.C.D.m - A.M.S (items/sec)	**184.8994**	184.4151	182.9559	**182.0317**
Normalized	100%	99.74%	98.95%	98.45%
Neural Magic DeepSparse - C.D.Y.C - A.M.S (ms/batch)	68.5103	**67.951**	**68.9309**	68.8771
Normalized	99.18%	100%	98.58%	98.66%
Neural Magic DeepSparse - C.D.Y.C - A.M.S (items/sec)	116.6551	**117.6106**	**115.9903**	116.0617
Normalized	99.19%	100%	98.62%	98.68%
Neural Magic DeepSparse - C.D.Y.C - S.S.S (ms/batch)	**11.2523**	11.2464	**11.2104**	11.2263
Normalized	99.63%	99.68%	100%	99.86%
Neural Magic DeepSparse - C.D.Y.C - S.S.S (items/sec)	**88.7219**	88.7951	**89.0528**	88.9476
Normalized	99.63%	99.71%	100%	99.88%
Neural Magic DeepSparse - C.C.R.5.I - A.M.S (ms/batch)	26.8102	26.8037	**26.7761**	**26.8519**
Normalized	99.87%	99.9%	100%	99.72%

deepsparse 7950x

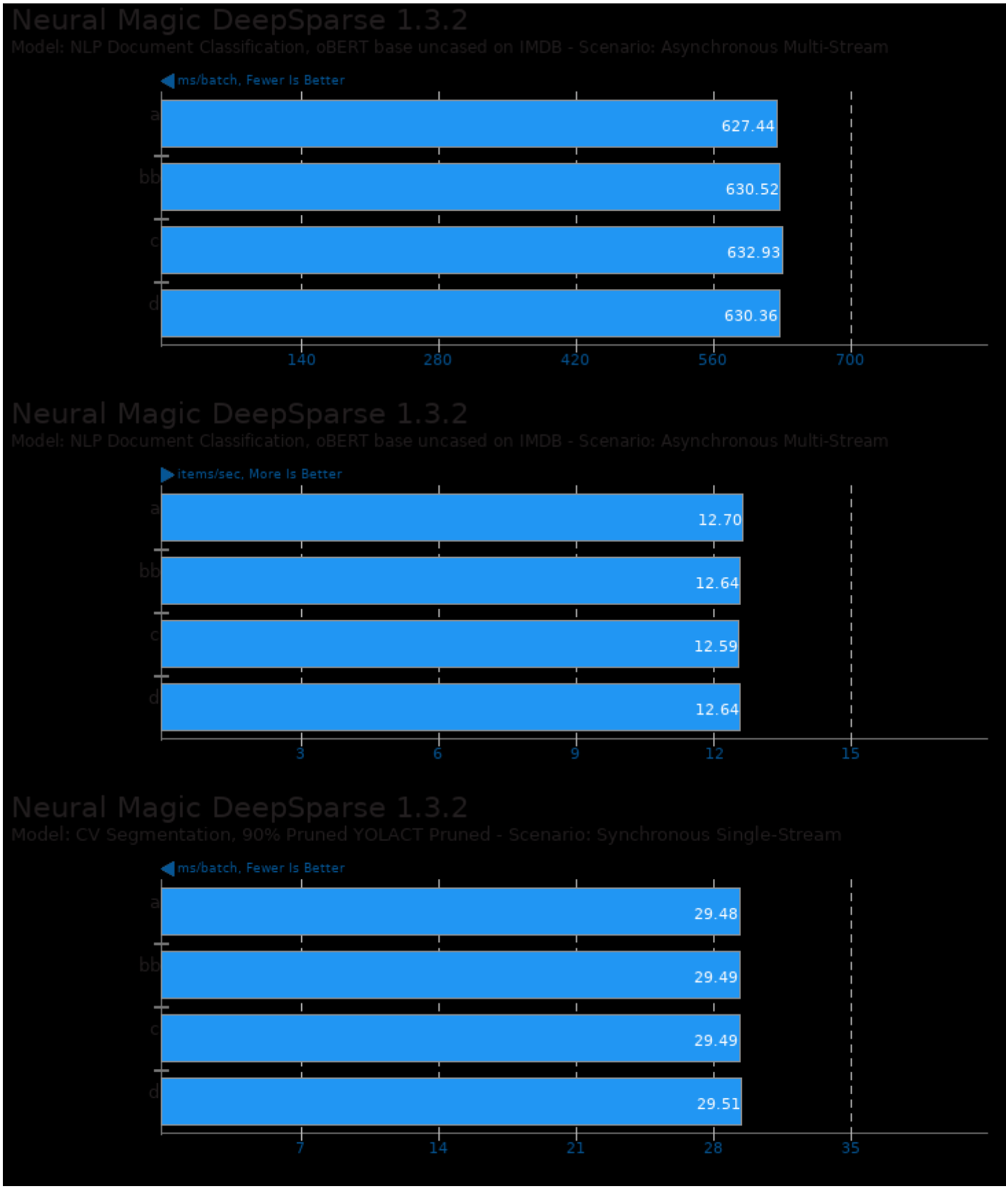
Neural Magic DeepSparse - C.C.R.5.I - A.M.S (items/sec)	298.2713	298.3351	**298.6177**	**297.798**
Normalized	99.88%	99.91%	100%	99.73%
Neural Magic DeepSparse - C.C.R.5.I - S.S.S (ms/batch)	4.888	**4.896**	**4.8812**	4.8865
Normalized	99.86%	99.7%	100%	99.89%
Neural Magic DeepSparse - C.C.R.5.I - S.S.S (items/sec)	204.3664	**204.0369**	**204.6633**	204.44
Normalized	99.85%	99.69%	100%	99.89%


Neural Magic DeepSparse 1.3.2
Model: NLP Token Classification, BERT base uncased conll2003 - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
a
633.18
bb
630.80
c
631.30
d
632.16
140
280
420
560
700
Neural Magic DeepSparse 1.3.2
Model: NLP Token Classification, BERT base uncased conll2003 - Scenario: Asynchronous Multi-Stream
items/sec, More Is Better
a
12.59
bb
12.63
c
12.63
d
12.59
3
6
9
12
15
Neural Magic DeepSparse 1.3.2
Model: CV Segmentation, 90% Pruned YOLACT Pruned - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
a
201.01
bb
201.36
c
200.48
d
200.53
40
80
120
160
200

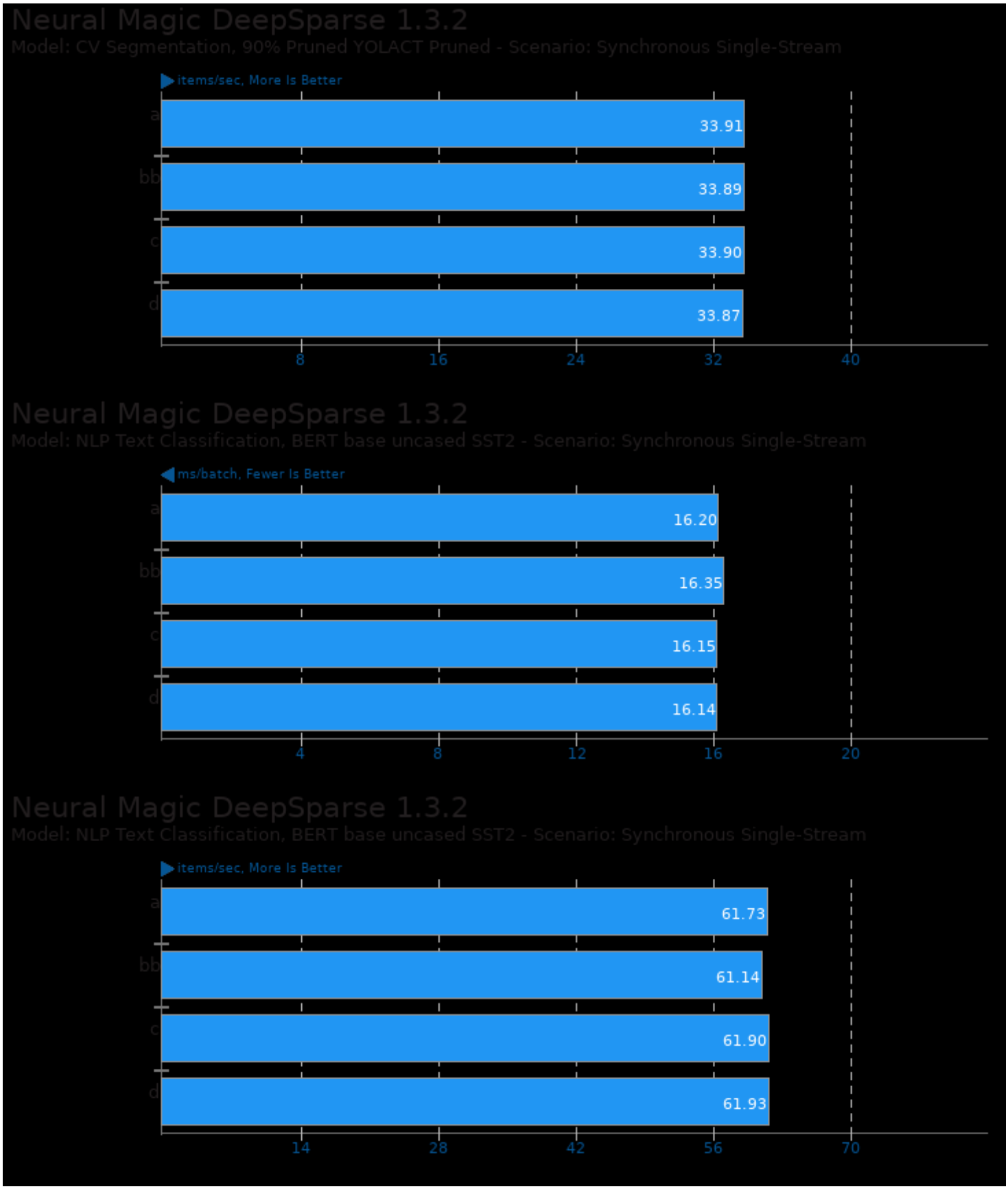
deepsparse 7950x

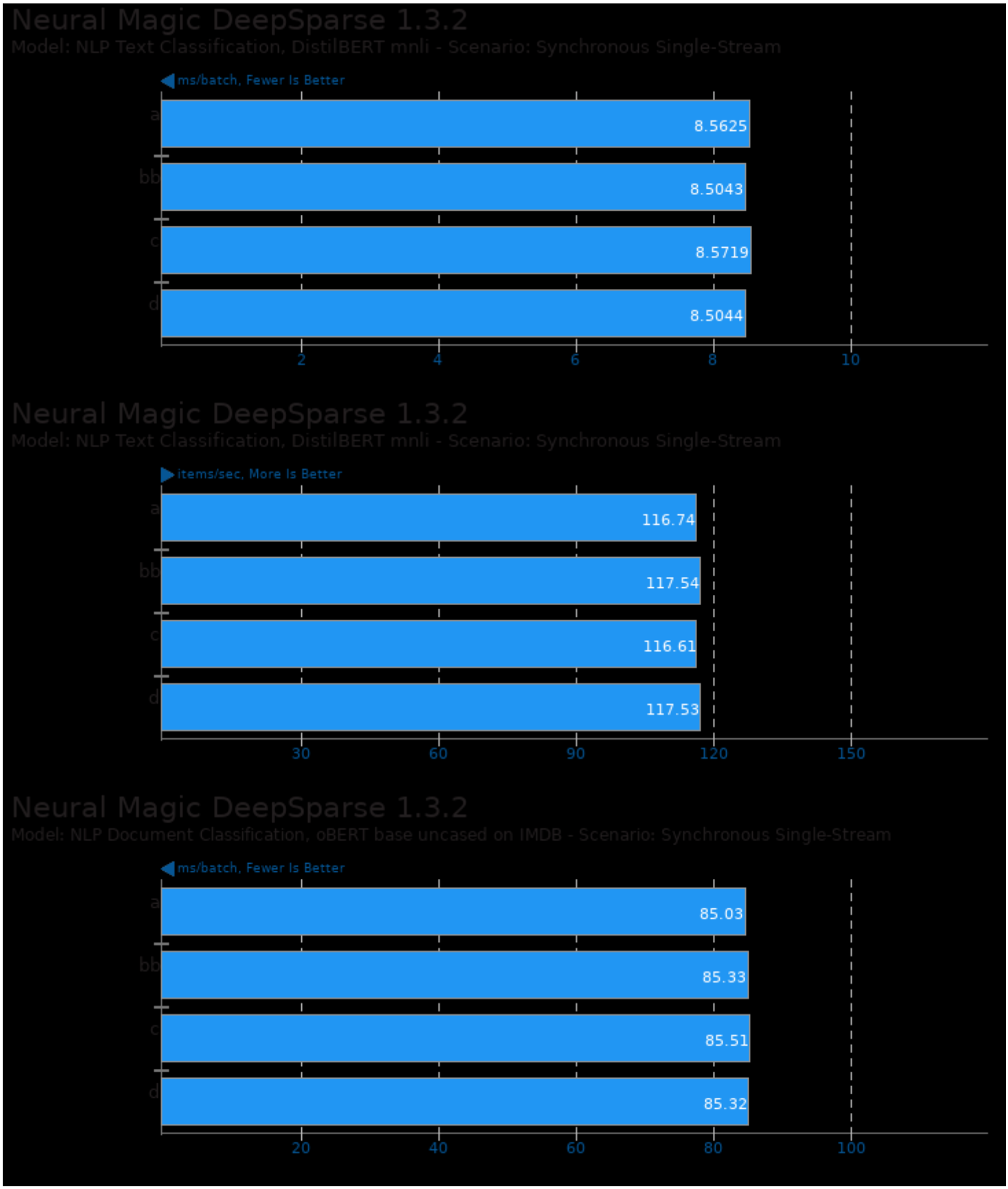


deepsparse 7950x

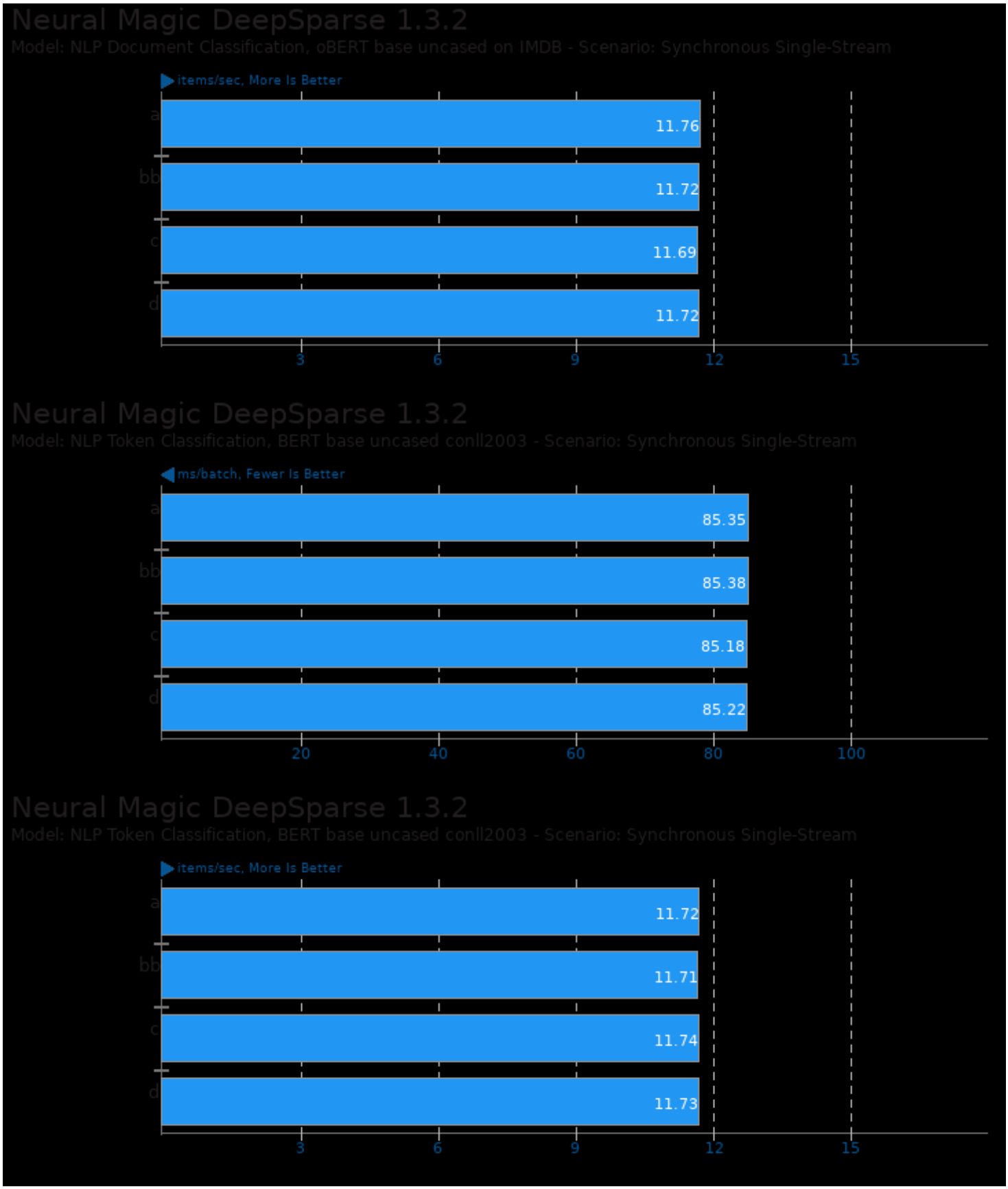


deepsparse 7950x

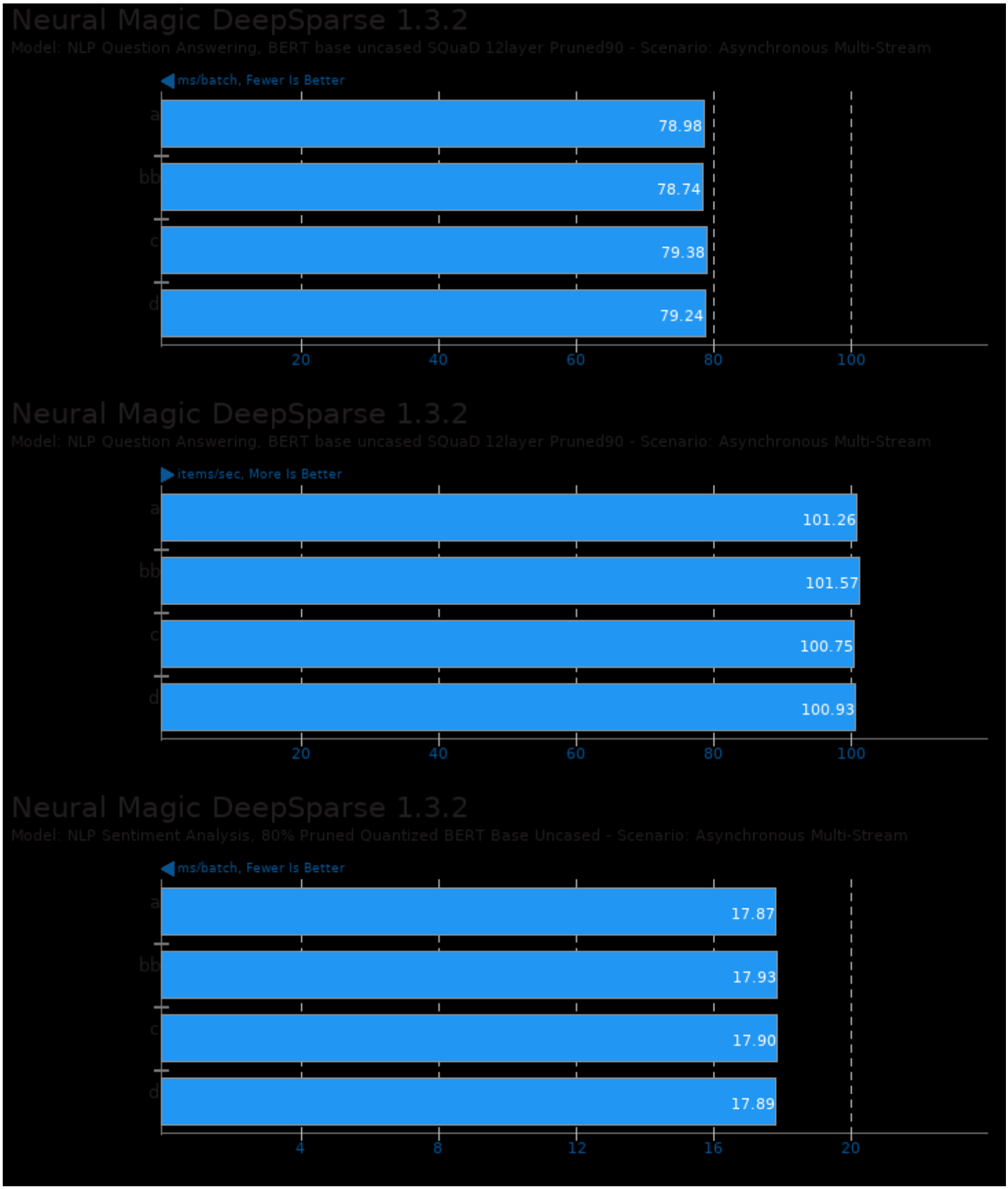



Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, DistilBERT mnli - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
a
8.5625
bb
8.5043
c
8.5719
d
8.5044
2
4
6
8
10
Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, DistilBERT mnli - Scenario: Synchronous Single-Stream
items/sec, More Is Better
a
116.74
bb
117.54
c
116.61
d
117.53
30
60
90
120
150
Neural Magic DeepSparse 1.3.2
Model: NLP Document Classification, oBERT base uncased on IMDB - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
a
85.03
bb
85.33
c
85.51
d
85.32
20
40
60
80
100

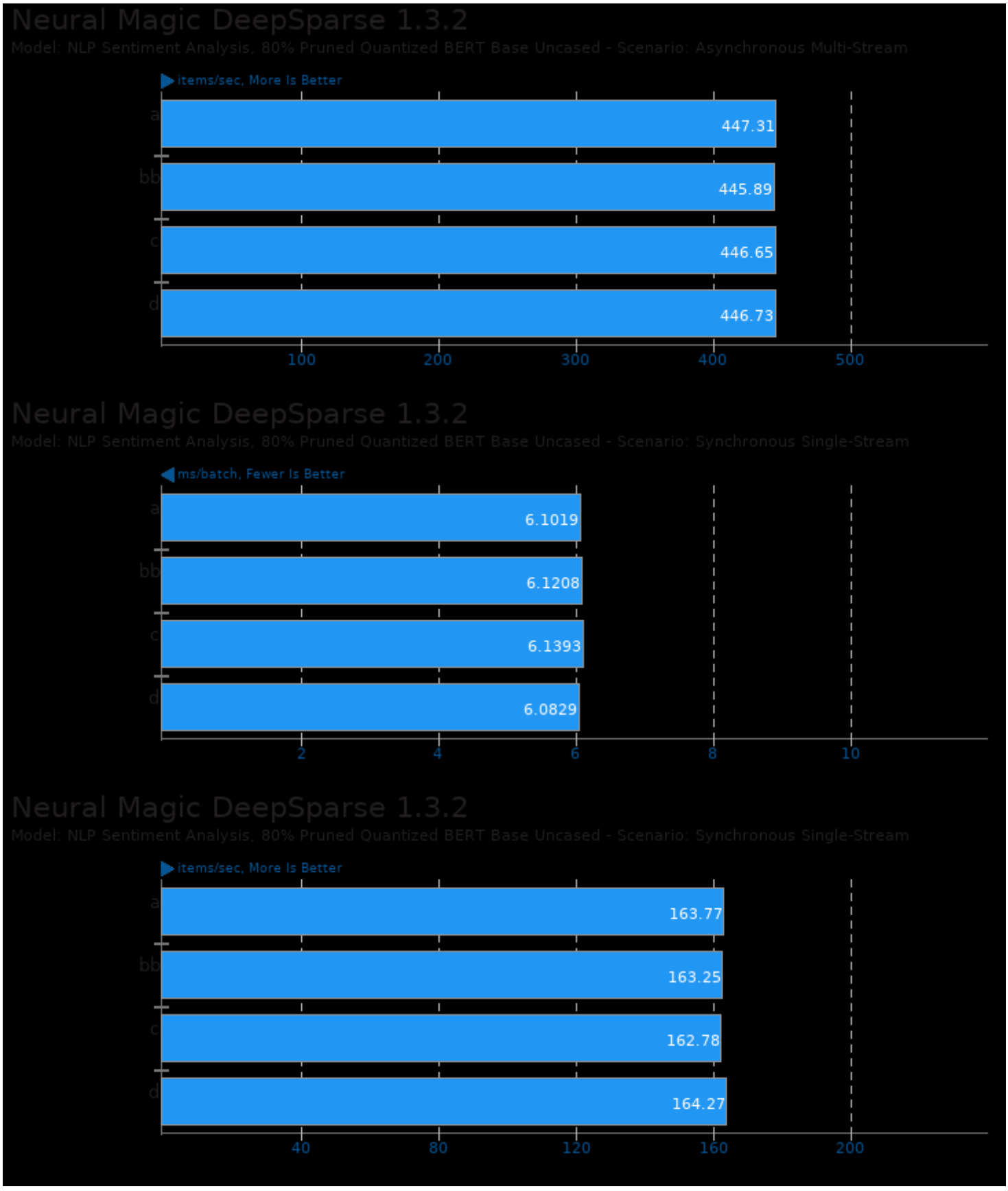
deepsparse 7950x



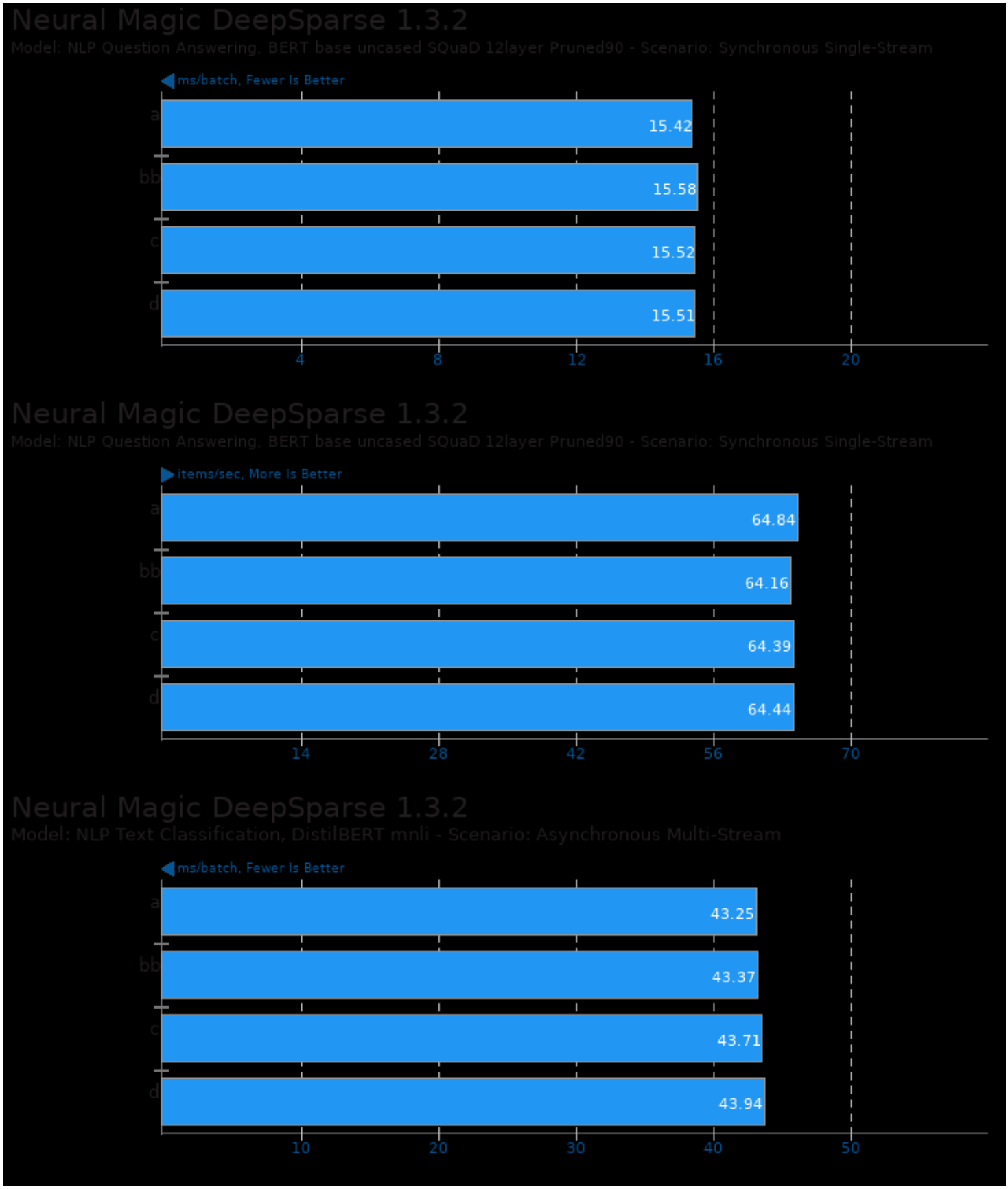
deepsparse 7950x



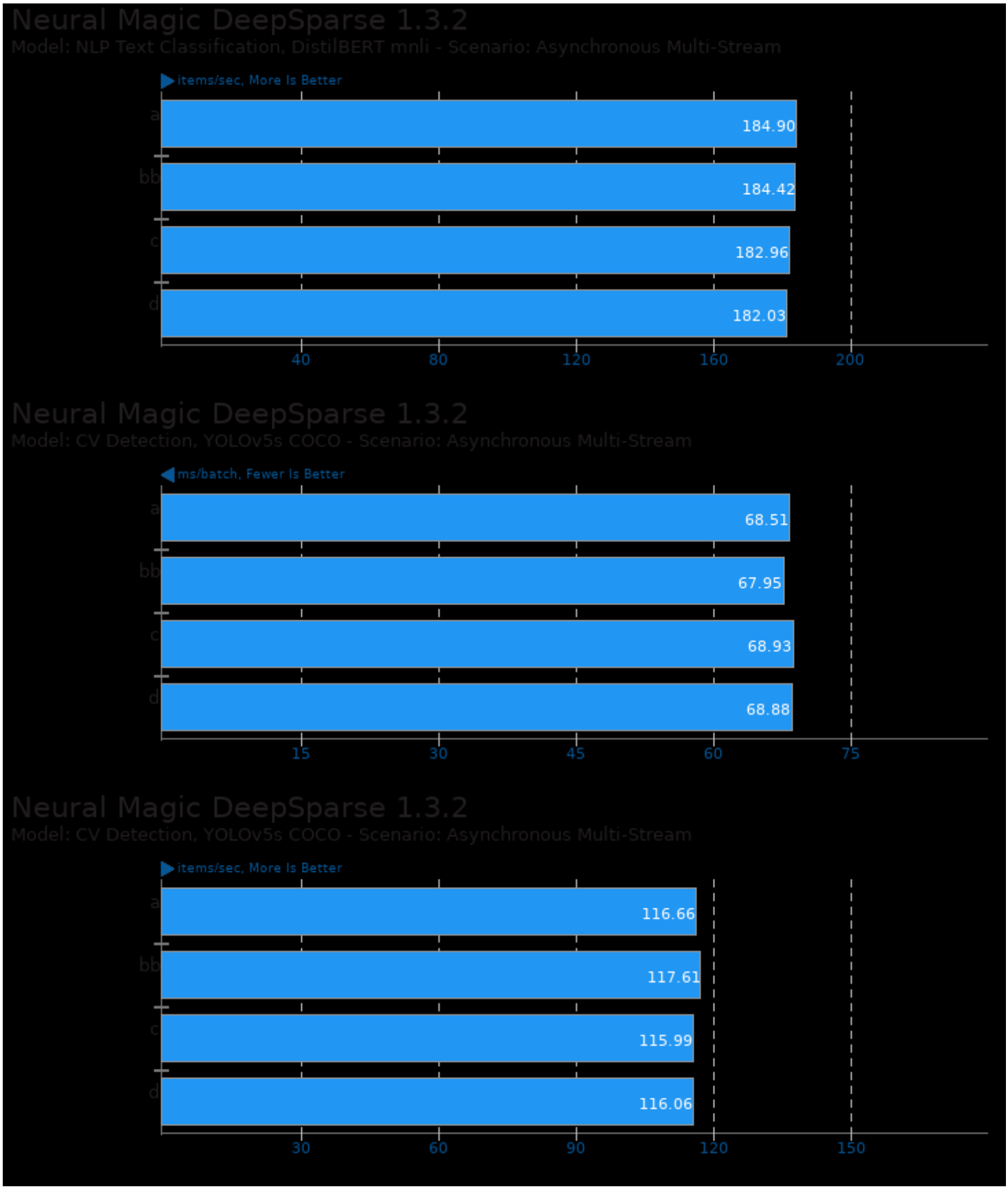
deepsparse 7950x



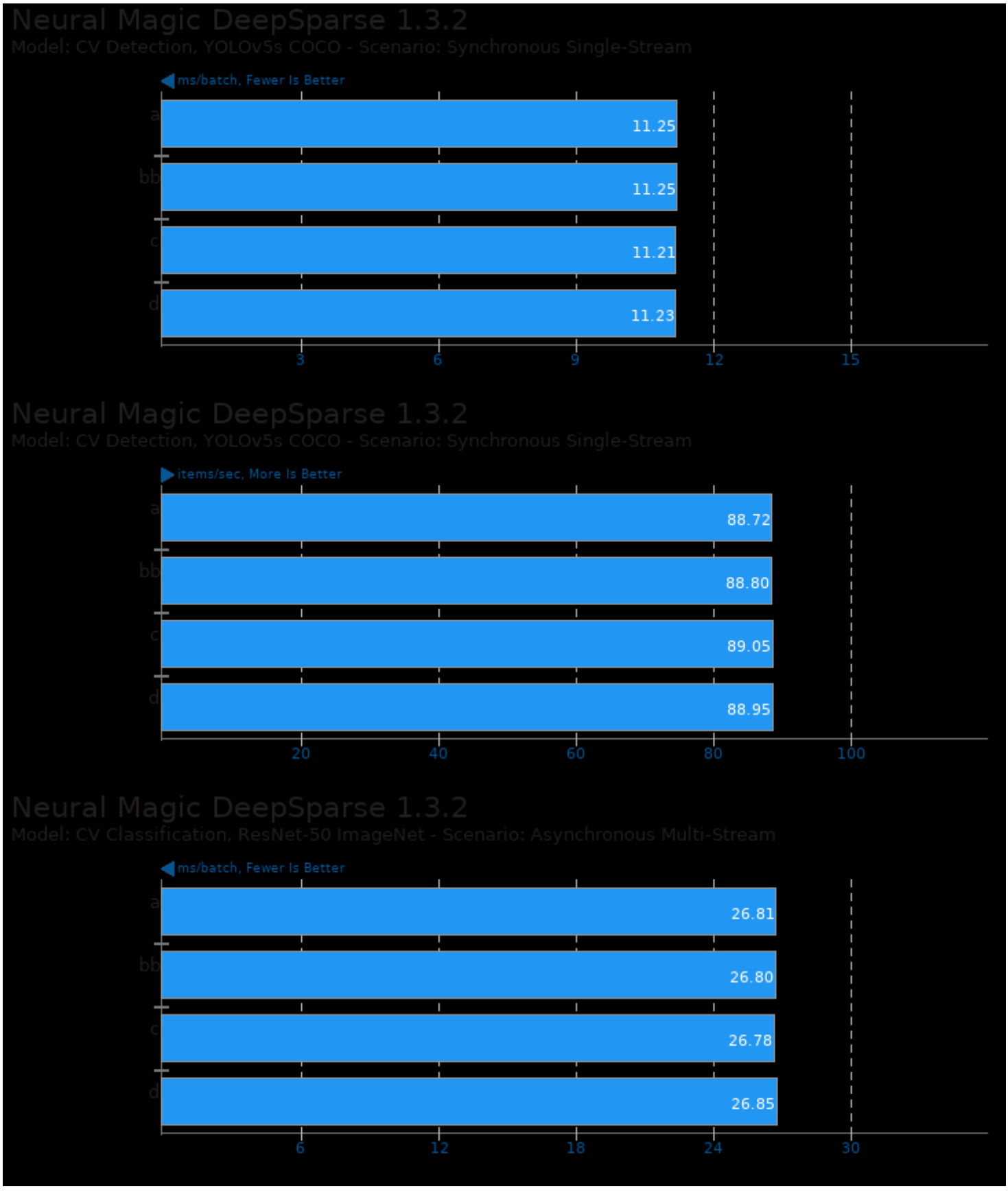
deepsparse 7950x

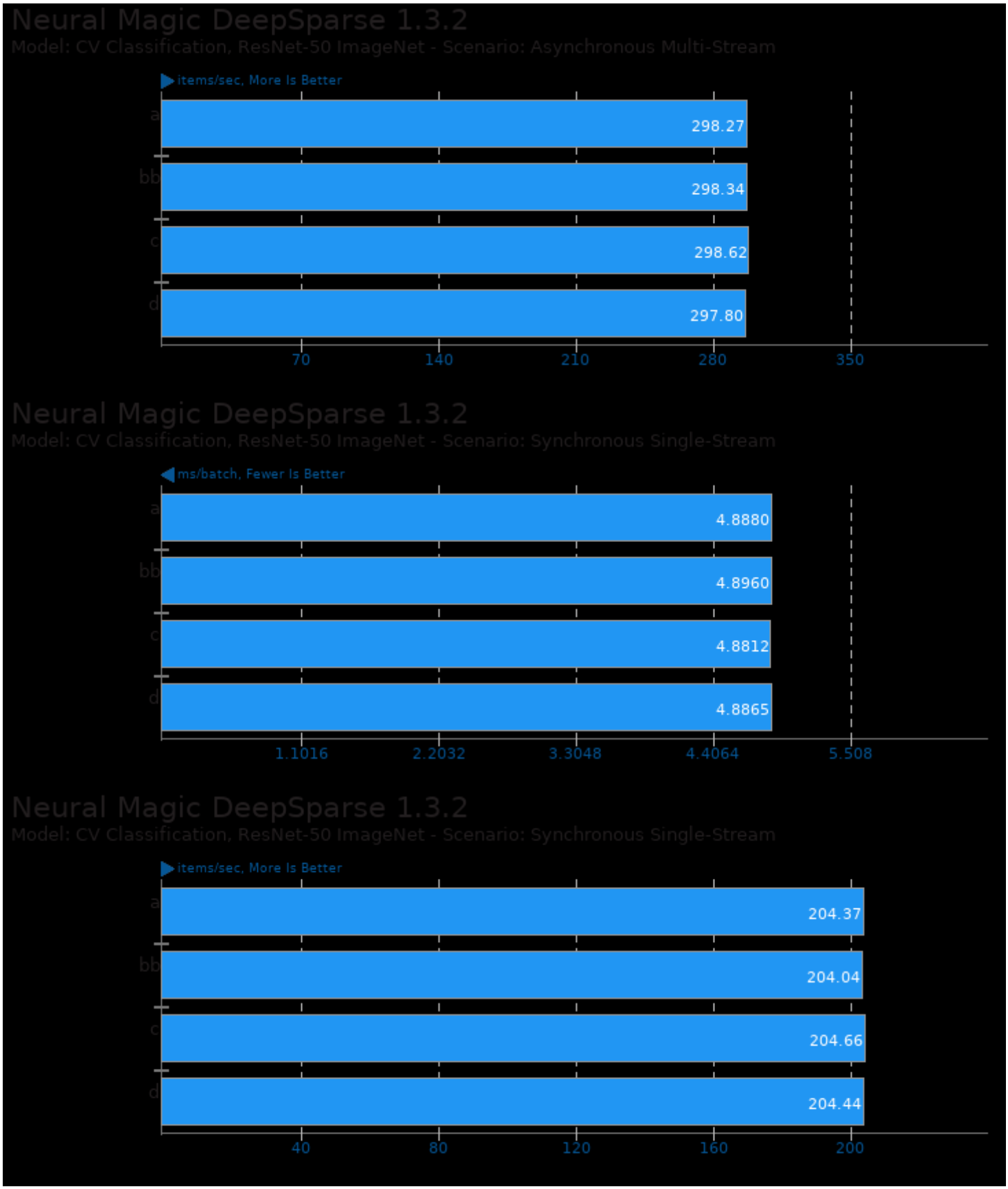


deepsparse 7950x



deepsparse 7950x





This file was automatically generated via the Phoronix Test Suite benchmarking software on Thursday, 28 March 2024 15:16.