

www.phoronix-test-suite.com

deepsparse tr

AMD Ryzen Threadripper 3990X 64-Core testing with a Gigabyte TRX40 AORUS PRO WIFI (F4p BIOS) and AMD Radeon RX 5700 8GB on Ubuntu 22.10 via the Phoronix Test Suite.

Automated Executive Summary

c had the most wins, coming in first place for 47% of the tests.

Based on the geometric mean of all complete results, the fastest (c) was 1.01x the speed of the slowest (a). b was 0.998x the speed of c and a was 0.992x the speed of b.

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

Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Asynchronous Multi-Stream) at 1.05x
Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Asynchronous Multi-Stream) at 1.049x
Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Synchronous Single-Stream) at 1.048x
Neural Magic DeepSparse (Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Synchronous Single-Stream) at 1.048x
Neural Magic DeepSparse (Model: CV Detection, YOLOv5s COCO - Scenario: Synchronous Single-Stream) at 1.048x

Neural Magic DeepSparse (Model: CV Detection, YOLOv5s COCO - Scenario: Synchronous Single-Stream) at 1.048x
Neural Magic DeepSparse (Model: NLP Text Classification, DistilBERT mnli - Scenario: Synchronous Single-Stream) at 1.039x
Neural Magic DeepSparse (Model: NLP Text Classification, DistilBERT mnli - Scenario: Synchronous Single-Stream) at 1.039x
Neural Magic DeepSparse (Model: NLP Token Classification, BERT base uncased conll2003 - Scenario: Synchronous Single-Stream) at 1.034x
Neural Magic DeepSparse (Model: NLP Token Classification, BERT base uncased conll2003 - Scenario: Synchronous Single-Stream) at 1.034x.

Test Systems:

a

b

c

Processor: AMD Ryzen Threadripper 3990X 64-Core @ 2.90GHz (64 Cores / 128 Threads), Motherboard: Gigabyte TRX40 AORUS PRO WIFI (F4p BIOS), Chipset: AMD Starship/Matisse, Memory: 128GB, Disk: Samsung SSD 970 EVO Plus 500GB, Graphics: AMD Radeon RX 5700 8GB (1750/875MHz), Audio: AMD Navi 10 HDMI Audio, Monitor: DELL P2415Q, Network: Intel I211 + Intel Wi-Fi 6 AX200

OS: Ubuntu 22.10, Kernel: 6.1.0-rc8-phx-mglru (x86_64), Desktop: GNOME Shell 43.0, Display Server: X Server 1.21.1.4 + Wayland, OpenGL: 4.6 Mesa 22.2.1 (LLVM 15.0.2 DRM 3.49), Vulkan: 1.3.224, Compiler: GCC 12.2.0 + LLVM 15.0.2, File-System: ext4, Screen Resolution: 3840x2160

Kernel Notes: Transparent Huge Pages: madvise
Processor Notes: Scaling Governor: acpi-cpufreq schedutil (Boost: Enabled) - CPU Microcode: 0x8301055
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: Mitigation of untrained return thunk; SMT enabled with STIBP protection + 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 STIBP: always-on RSB filling PBRSB-eIBRS: Not affected + srbds: Not affected + tsx_async_abort: Not affected

	a	b	c
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - A.M.S (items/sec)	**34.8860**	34.4957	**34.2097**
Normalized	100%	98.88%	98.06%
Standard Deviation	1.9%		
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - S.S.S (items/sec)	13.6680	**13.6164**	**13.7707**
Normalized	99.25%	98.88%	100%
Standard Deviation	0.4%		

deepsparse tr

Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - A.M.S (items/sec)	**263.2698**	271.2283	**276.2048**
Normalized	95.32%	98.2%	100%
Standard Deviation	4%		
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - S.S.S (items/sec)	**47.4189**	**49.7013**	49.6628
Normalized	95.41%	100%	99.92%
Standard Deviation	1.8%		
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - A.M.S (items/sec)	**88.5450**	88.7699	**89.0653**
Normalized	99.42%	99.67%	100%
Standard Deviation	0.7%		
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - S.S.S (items/sec)	**20.8260**	21.1142	**21.2348**
Normalized	98.07%	99.43%	100%
Standard Deviation	1.5%		
Neural Magic DeepSparse - C.D.Y.C - A.M.S (items/sec)	171.8227	**171.5536**	**172.622**
Normalized	99.54%	99.38%	100%
Standard Deviation	0.1%		
Neural Magic DeepSparse - C.D.Y.C - S.S.S (items/sec)	**63.6711**	**60.7788**	62.2604
Normalized	100%	95.46%	97.78%
Standard Deviation	0.3%		
Neural Magic DeepSparse - C.C.R.5.I - A.M.S (items/sec)	**414.4296**	**415.6349**	414.552
Normalized	99.71%	100%	99.74%
Standard Deviation	0%		
Neural Magic DeepSparse - C.C.R.5.I - S.S.S (items/sec)	94.5820	**95.0838**	**93.6235**
Normalized	99.47%	100%	98.46%
Standard Deviation	0.3%		
Neural Magic DeepSparse - N.T.C.D.m - A.M.S (items/sec)	**279.3454**	**284.1417**	283.3205
Normalized	98.31%	100%	99.71%
Standard Deviation	0.4%		
Neural Magic DeepSparse - N.T.C.D.m - S.S.S (items/sec)	**57.8622**	**60.1136**	59.4981
Normalized	96.25%	100%	98.98%
Standard Deviation	1%		
Neural Magic DeepSparse - C.S.9.P.Y.P - A.M.S (items/sec)	**43.5414**	**43.8726**	43.7343
Normalized	99.25%	100%	99.68%
Standard Deviation	0.2%		
Neural Magic DeepSparse - C.S.9.P.Y.P - S.S.S (items/sec)	**18.3022**	18.3262	**18.3593**
Normalized	99.69%	99.82%	100%
Standard Deviation	0.1%		
Neural Magic DeepSparse - N.T.C.B.b.u.S - A.M.S (items/sec)	**142.2132**	**143.9332**	143.714
Normalized	98.81%	100%	99.85%
Standard Deviation	0%		

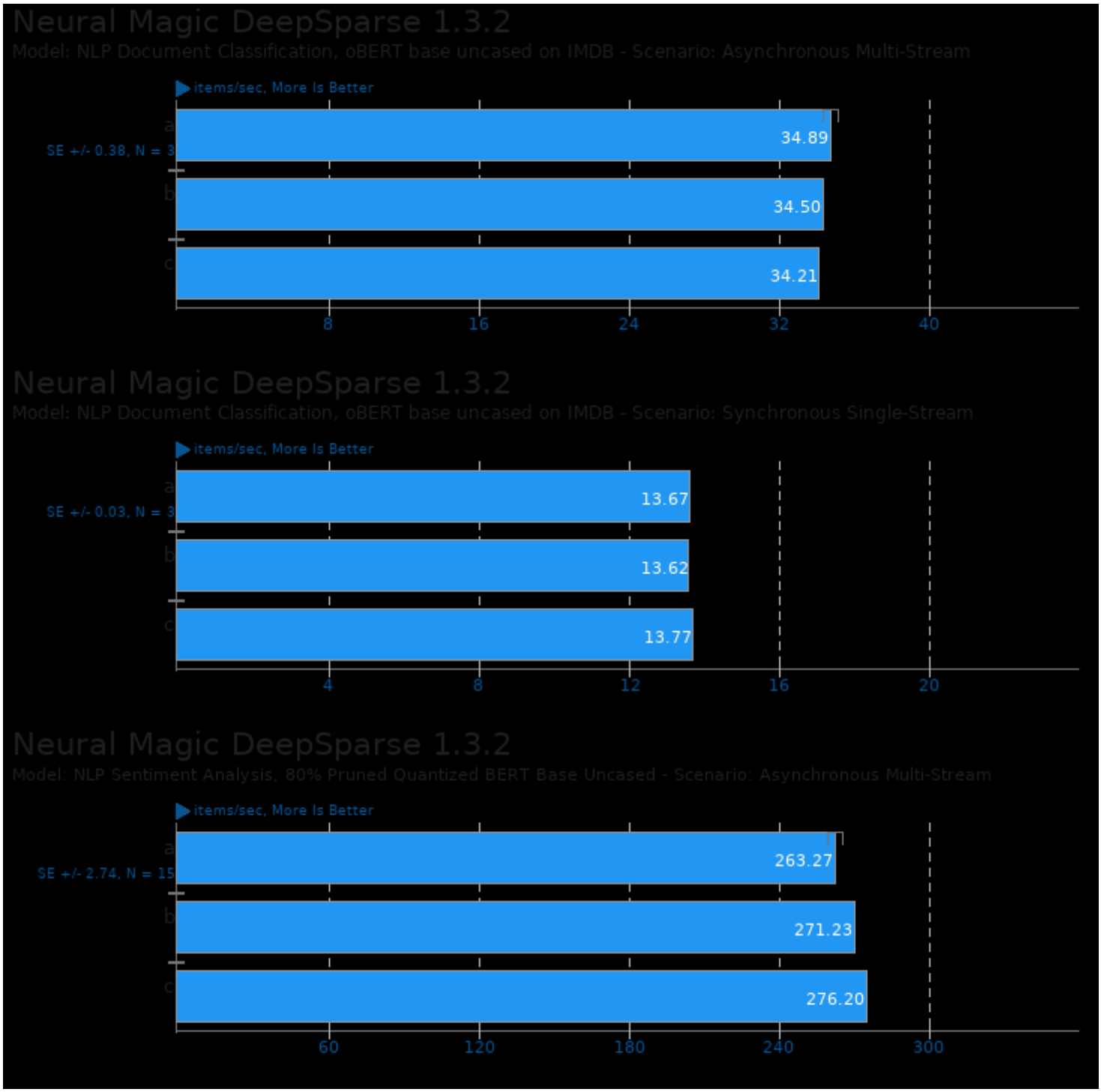
deepsparse tr

Neural Magic DeepSparse - N.T.C.B.b.u.S - S.S.S (items/sec)	30.5467	**30.4182**	**31.0025**
Normalized	98.53%	98.12%	100%
Standard Deviation	0.7%		
Neural Magic DeepSparse - N.T.C.B.b.u.c - A.M.S (items/sec)	**33.4855**	33.6307	**33.6845**
Normalized	99.41%	99.84%	100%
Standard Deviation	0.3%		
Neural Magic DeepSparse - N.T.C.B.b.u.c - S.S.S (items/sec)	**13.4049**	**13.8574**	13.7254
Normalized	96.73%	100%	99.05%
Standard Deviation	0.2%		
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - A.M.S (ms/batch)	**911.3941**	916.5997	**929.8296**
Normalized	100%	99.43%	98.02%
Standard Deviation	1.5%		
Neural Magic DeepSparse - N.D.C.o.b.u.o.I - S.S.S (ms/batch)	73.1554	**73.4315**	**72.6091**
Normalized	99.25%	98.88%	100%
Standard Deviation	0.4%		
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - A.M.S (ms/batch)	**121.5987**	117.8674	**115.7776**
Normalized	95.21%	98.23%	100%
Standard Deviation	4%		
Neural Magic DeepSparse - N.S.A.8.P.Q.B.B.U - S.S.S (ms/batch)	**21.0843**	**20.1109**	20.1262
Normalized	95.38%	100%	99.92%
Standard Deviation	1.8%		
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - A.M.S (ms/batch)	**360.5545**	359.2449	**358.0307**
Normalized	99.3%	99.66%	100%
Standard Deviation	0.8%		
Neural Magic DeepSparse - N.Q.A.B.b.u.S.1.P - S.S.S (ms/batch)	**48.0131**	47.3506	**47.0812**
Normalized	98.06%	99.43%	100%
Standard Deviation	1.5%		
Neural Magic DeepSparse - C.D.Y.C - A.M.S (ms/batch)	185.7590	**185.9302**	**184.7932**
Normalized	99.48%	99.39%	100%
Standard Deviation	0.1%		
Neural Magic DeepSparse - C.D.Y.C - S.S.S (ms/batch)	**15.6921**	**16.4399**	16.0481
Normalized	100%	95.45%	97.78%
Standard Deviation	0.3%		
Neural Magic DeepSparse - C.C.R.5.I - A.M.S (ms/batch)	**77.1694**	**76.9125**	77.1478
Normalized	99.67%	100%	99.7%
Standard Deviation	0%		
Neural Magic DeepSparse - C.C.R.5.I - S.S.S (ms/batch)	10.5637	**10.5084**	**10.6724**
Normalized	99.48%	100%	98.46%
Standard Deviation	0.3%		

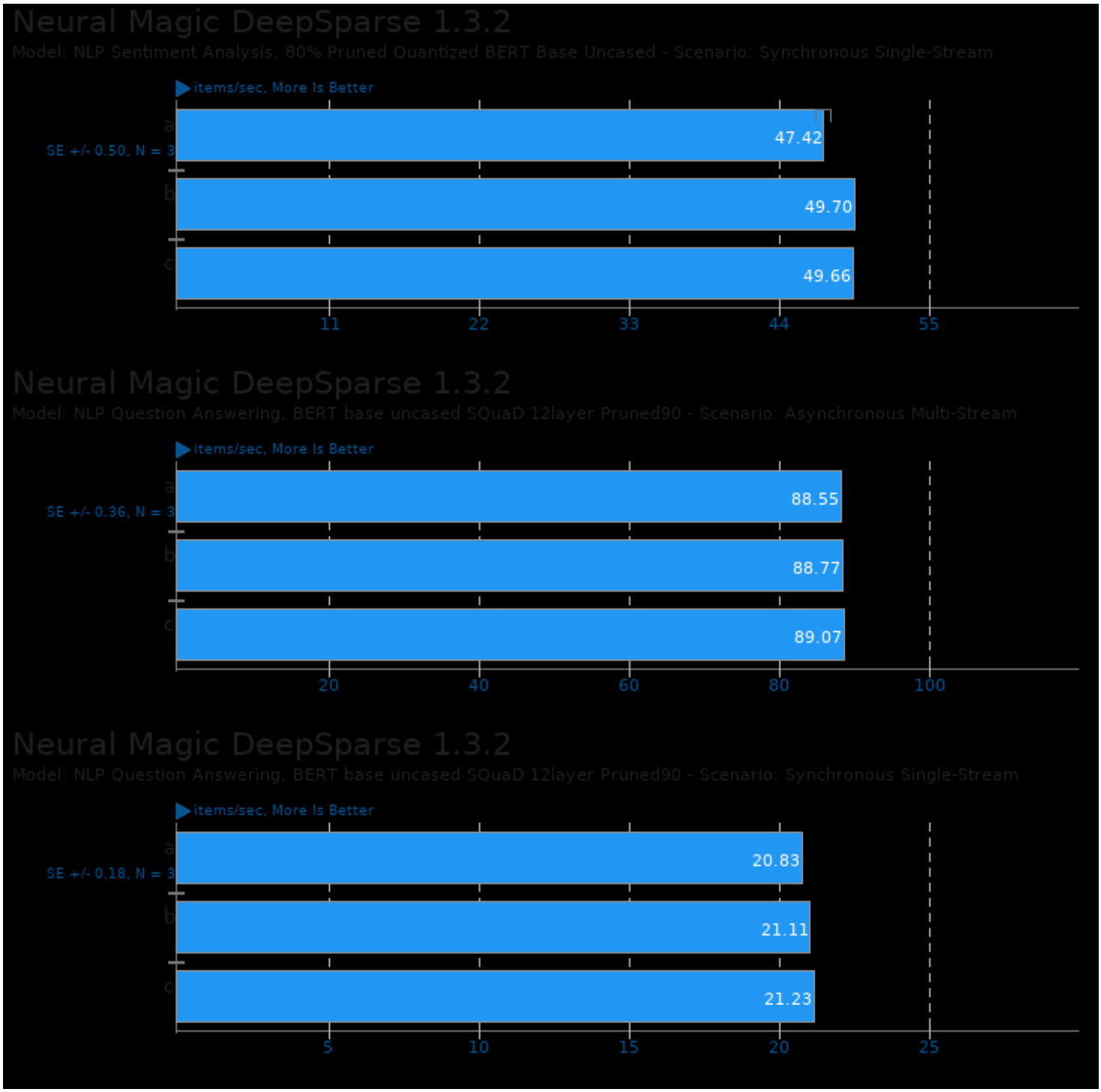
deepsparse tr

Neural Magic DeepSparse - N.T.C.D.m - A.M.S (ms/batch)	**114.4637**	**112.5672**	112.8945
Normalized	98.34%	100%	99.71%
Standard Deviation	0.3%		
Neural Magic DeepSparse - N.T.C.D.m - S.S.S (ms/batch)	**17.2750**	**16.627**	16.7985
Normalized	96.25%	100%	98.98%
Standard Deviation	1%		
Neural Magic DeepSparse - C.S.9.P.Y.P - A.M.S (ms/batch)	**728.3385**	725.2629	**724.5184**
Normalized	99.48%	99.9%	100%
Standard Deviation	0.1%		
Neural Magic DeepSparse - C.S.9.P.Y.P - S.S.S (ms/batch)	**54.6204**	54.5493	**54.4505**
Normalized	99.69%	99.82%	100%
Standard Deviation	0.1%		
Neural Magic DeepSparse - N.T.C.B.b.u.S - A.M.S (ms/batch)	**224.9088**	**222.2642**	222.6035
Normalized	98.82%	100%	99.85%
Standard Deviation	0%		
Neural Magic DeepSparse - N.T.C.B.b.u.S - S.S.S (ms/batch)	32.7294	**32.8655**	**32.2466**
Normalized	98.52%	98.12%	100%
Standard Deviation	0.7%		
Neural Magic DeepSparse - N.T.C.B.b.u.c - A.M.S (ms/batch)	**943.8212**	941.5145	**937.8174**
Normalized	99.36%	99.61%	100%
Standard Deviation	0.2%		
Neural Magic DeepSparse - N.T.C.B.b.u.c - S.S.S (ms/batch)	**74.5910**	**72.1548**	72.8486
Normalized	96.73%	100%	99.05%
Standard Deviation	0.2%		

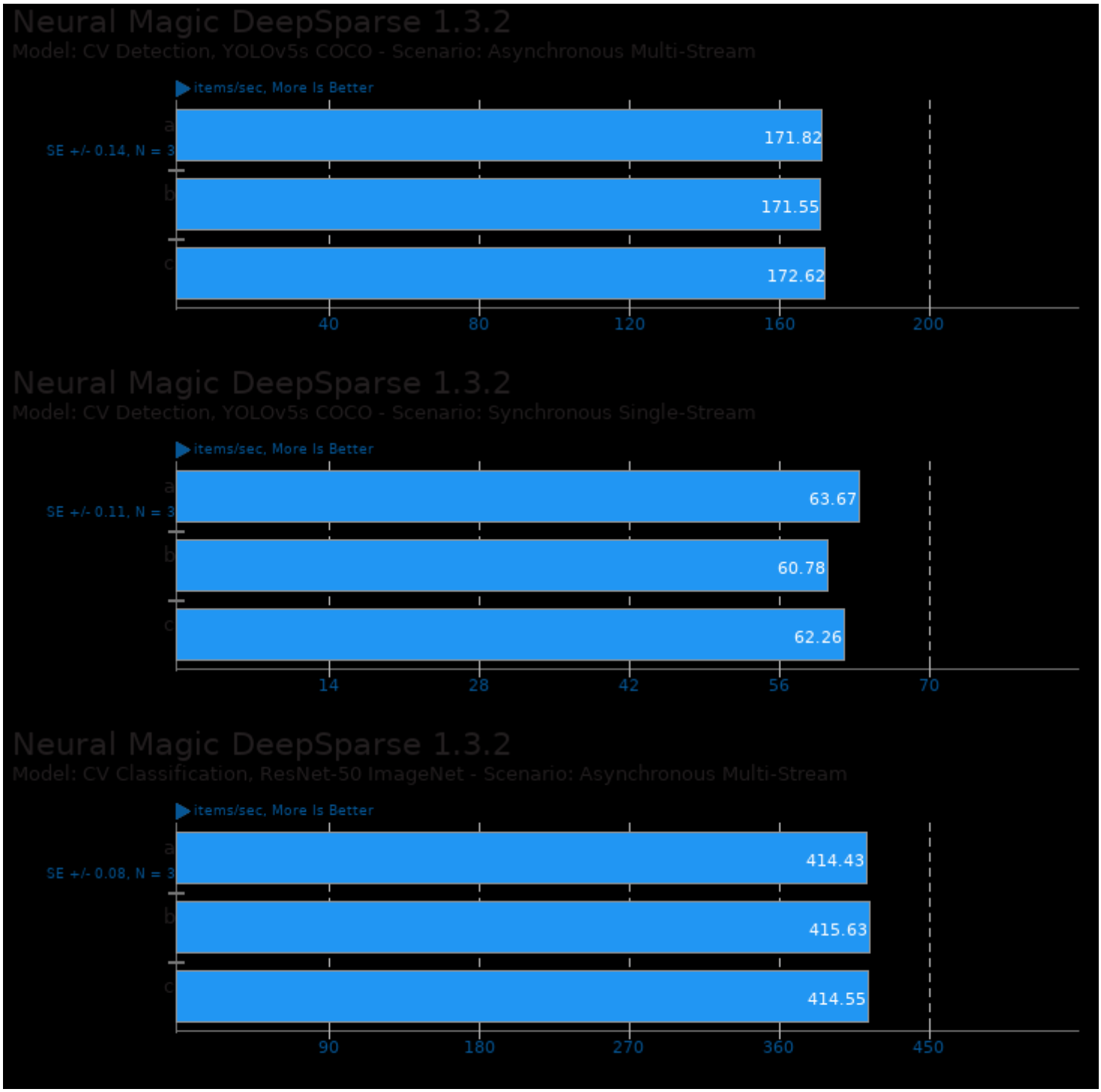
deepsparse tr



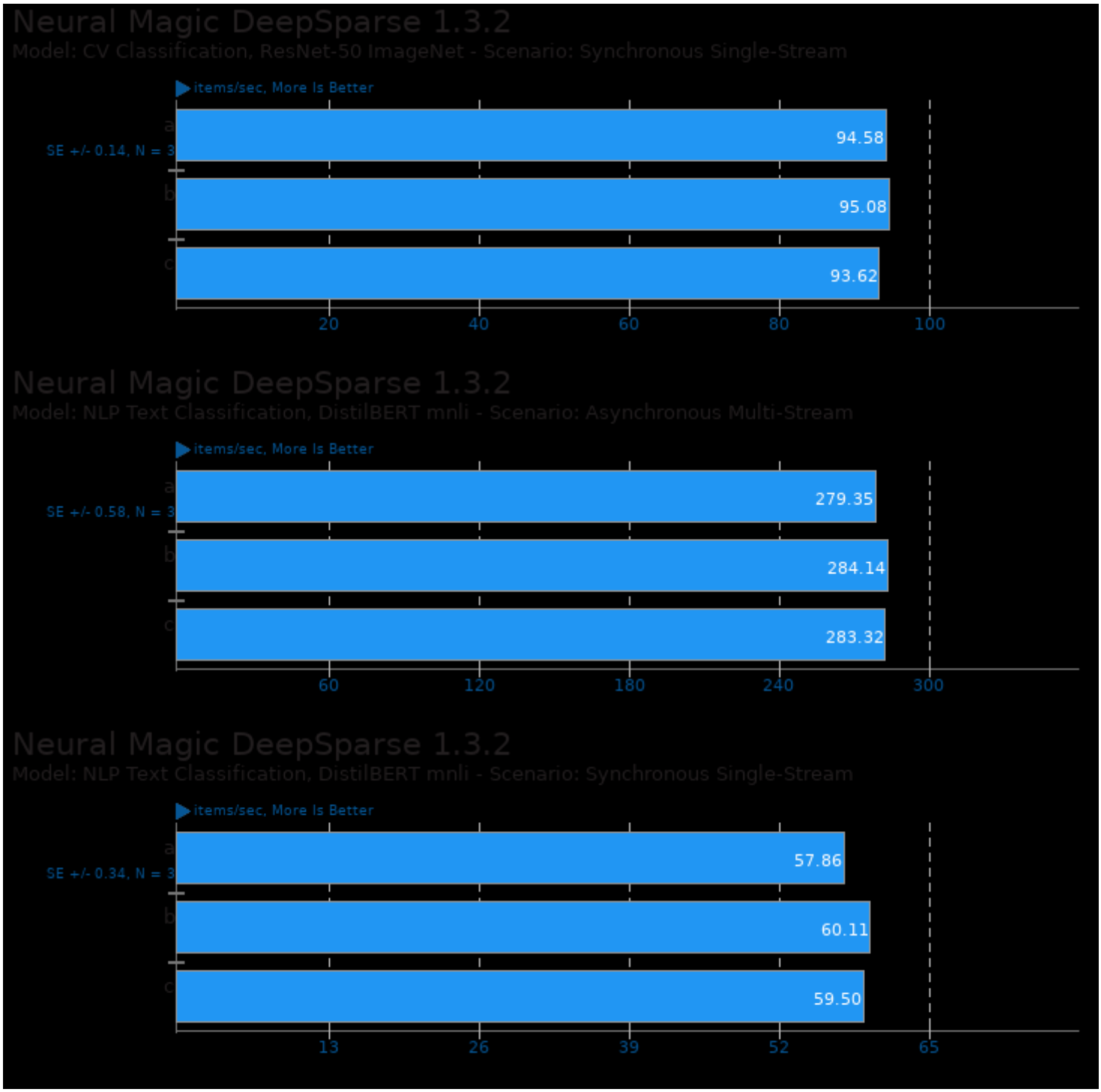
deepsparse tr

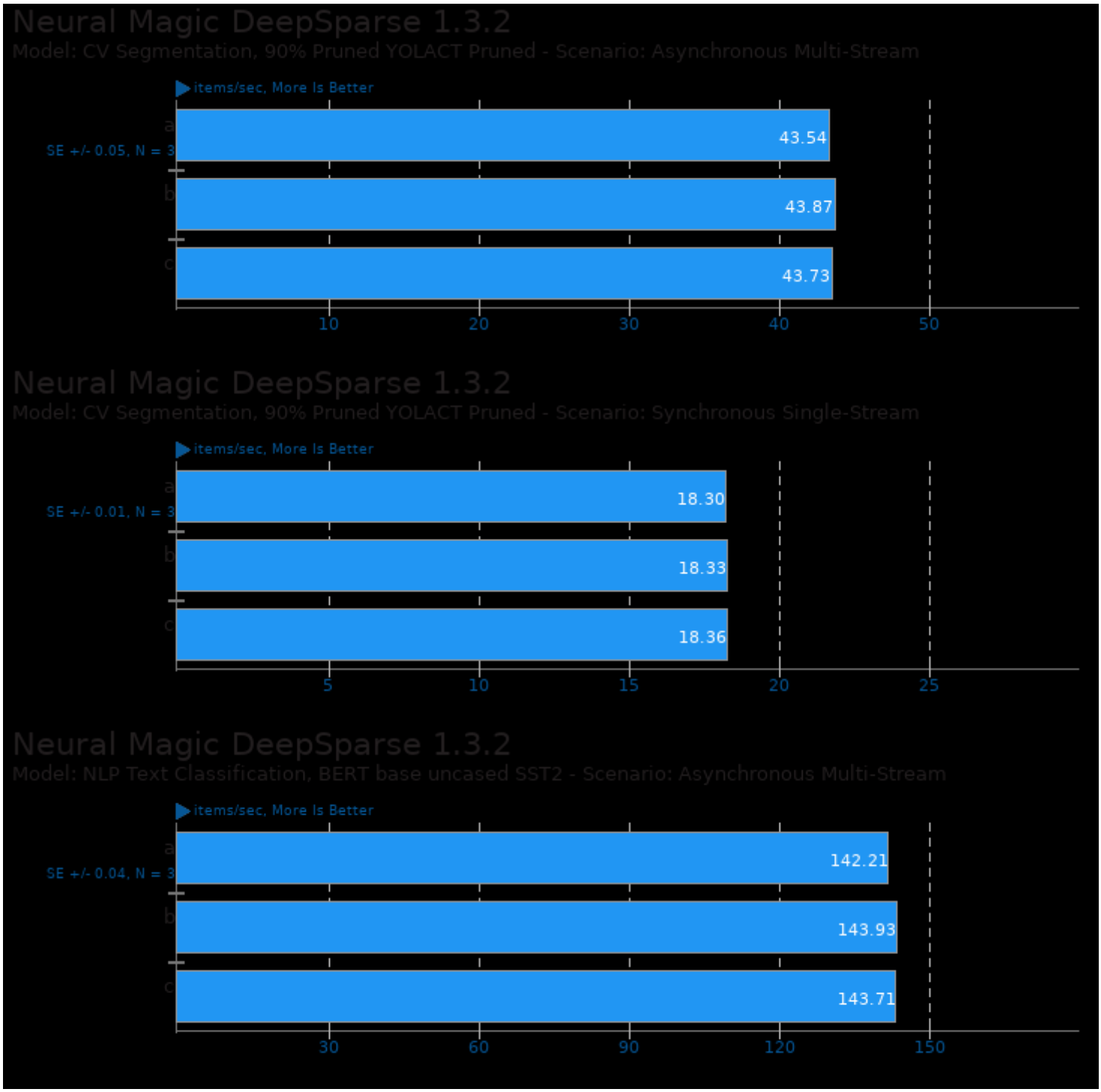


deepsparse tr

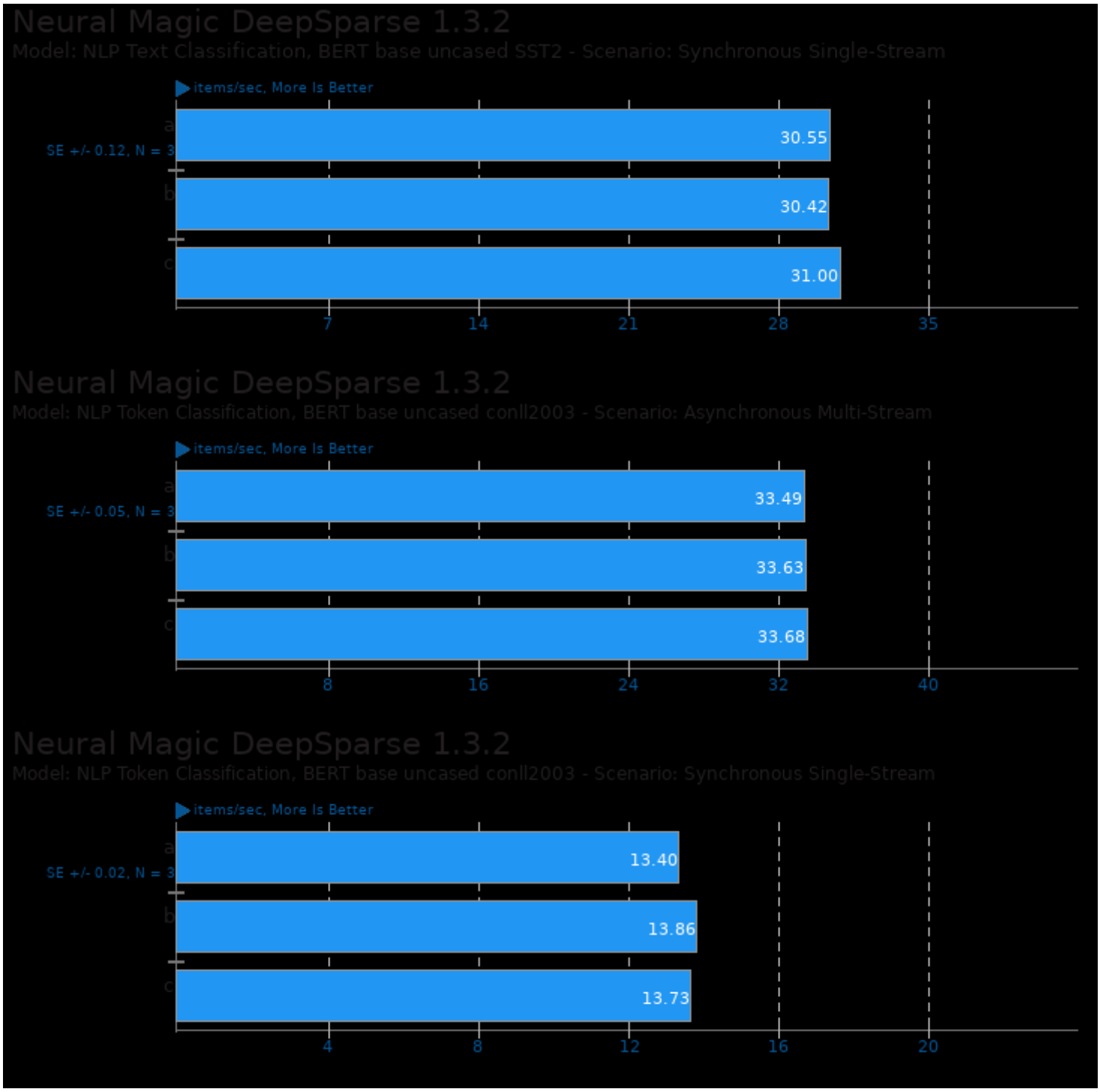


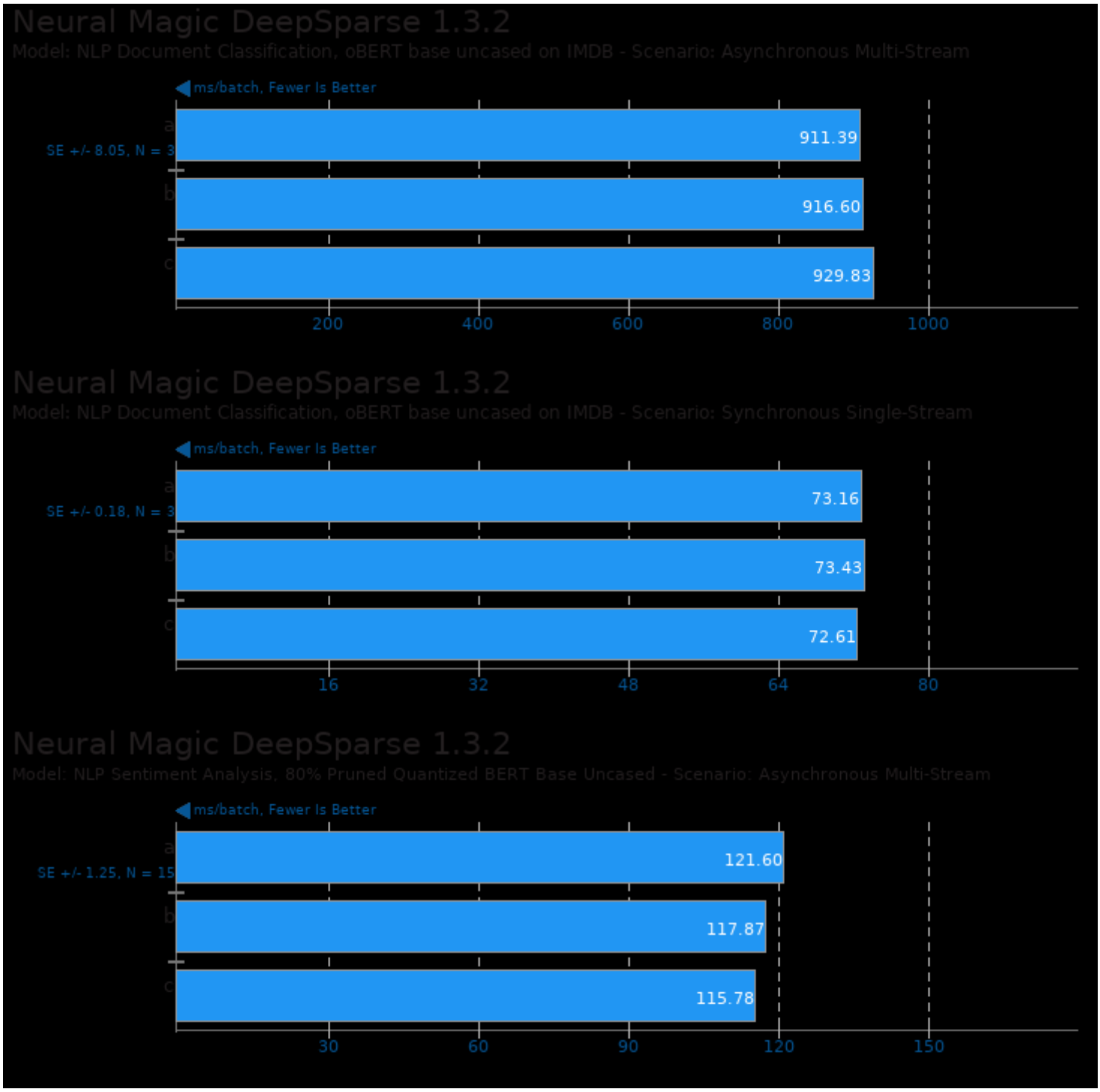
deepsparse tr

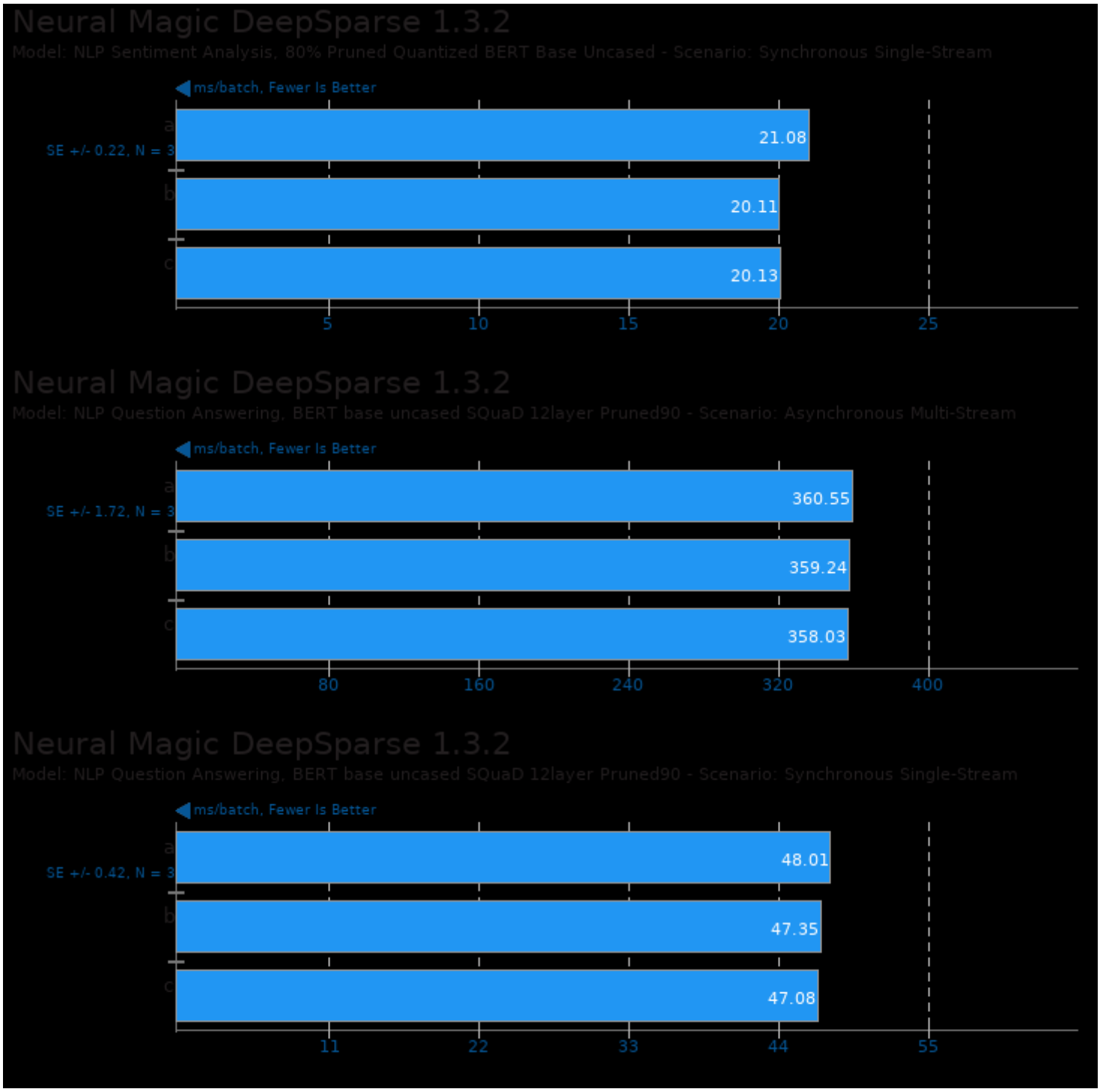



Neural Magic DeepSparse 1.3.2
Model: CV Segmentation, 90% Pruned YOLACT Pruned - Scenario: Asynchronous Multi-Stream
items/sec, More Is Better
SE +/- 0.05, N = 3
a
b
c
43.54
43.87
43.73
10
20
30
40
50
Neural Magic DeepSparse 1.3.2
Model: CV Segmentation, 90% Pruned YOLACT Pruned - Scenario: Synchronous Single-Stream
items/sec, More Is Better
SE +/- 0.01, N = 3
a
b
c
18.30
18.33
18.36
5
10
15
20
25
Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, BERT base uncased SST2 - Scenario: Asynchronous Multi-Stream
items/sec, More Is Better
SE +/- 0.04, N = 3
a
b
c
142.21
143.93
143.71
30
60
90
120
150

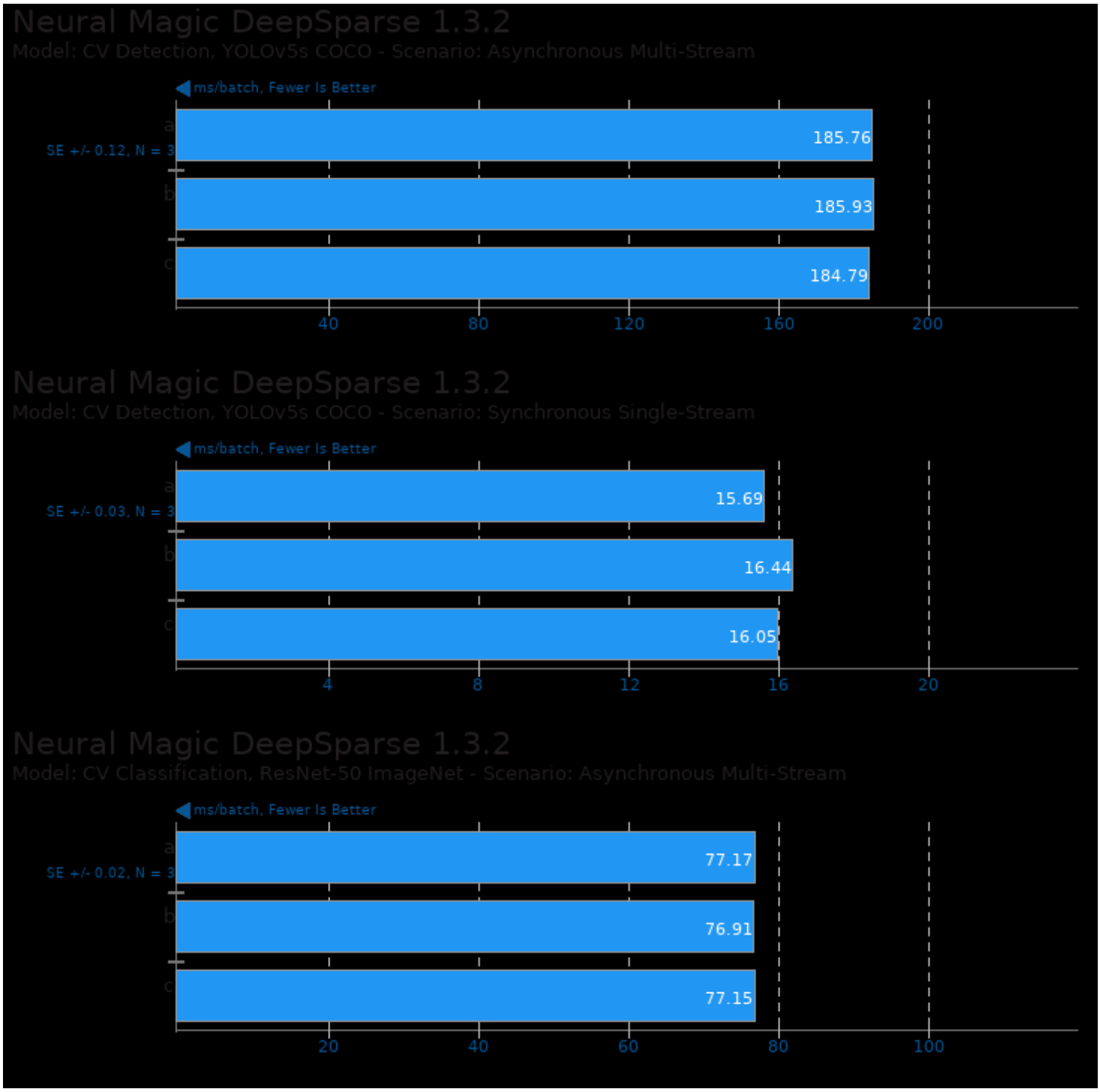
deepsparse tr

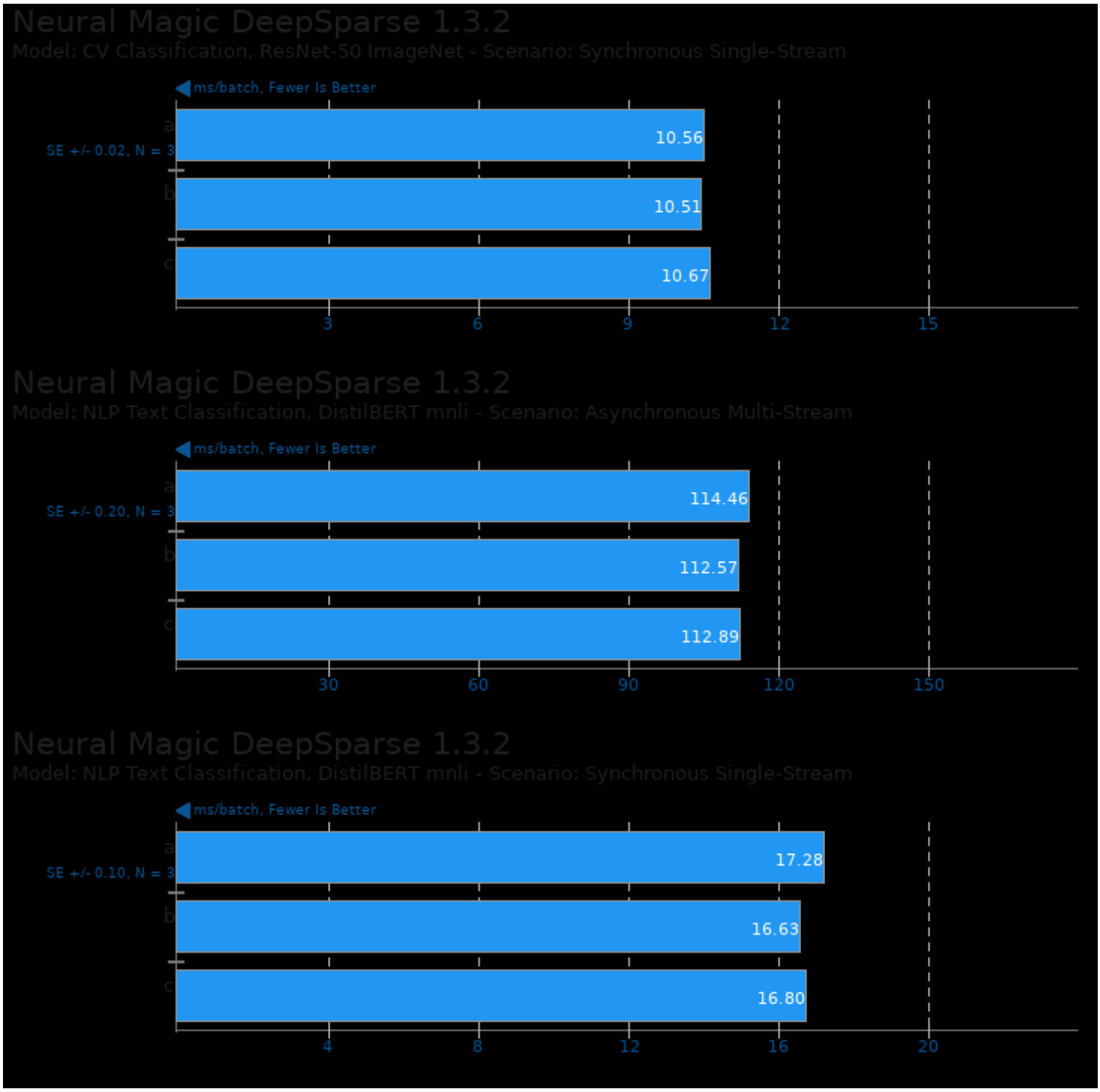


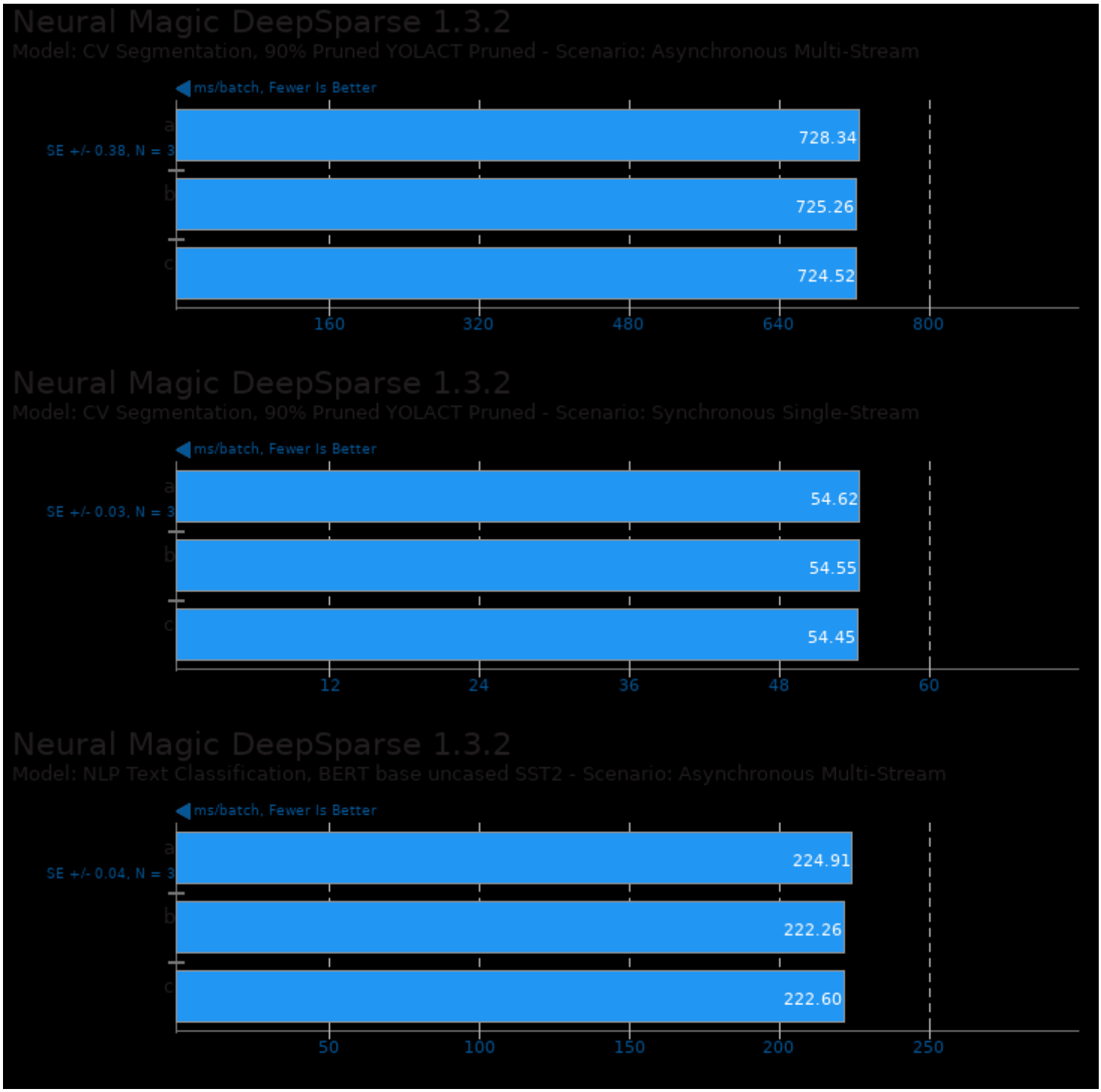

Neural Magic DeepSparse 1.3.2
Model: NLP Document Classification, oBERT base uncased on IMDB - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 8.05, N = 3
911.39
916.60
929.83
200
400
600
800
1000
Neural Magic DeepSparse 1.3.2
Model: NLP Document Classification, oBERT base uncased on IMDB - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.18, N = 3
73.16
73.43
72.61
16
32
48
64
80
Neural Magic DeepSparse 1.3.2
Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 1.25, N = 15
121.60
117.87
115.78
30
60
90
120
150

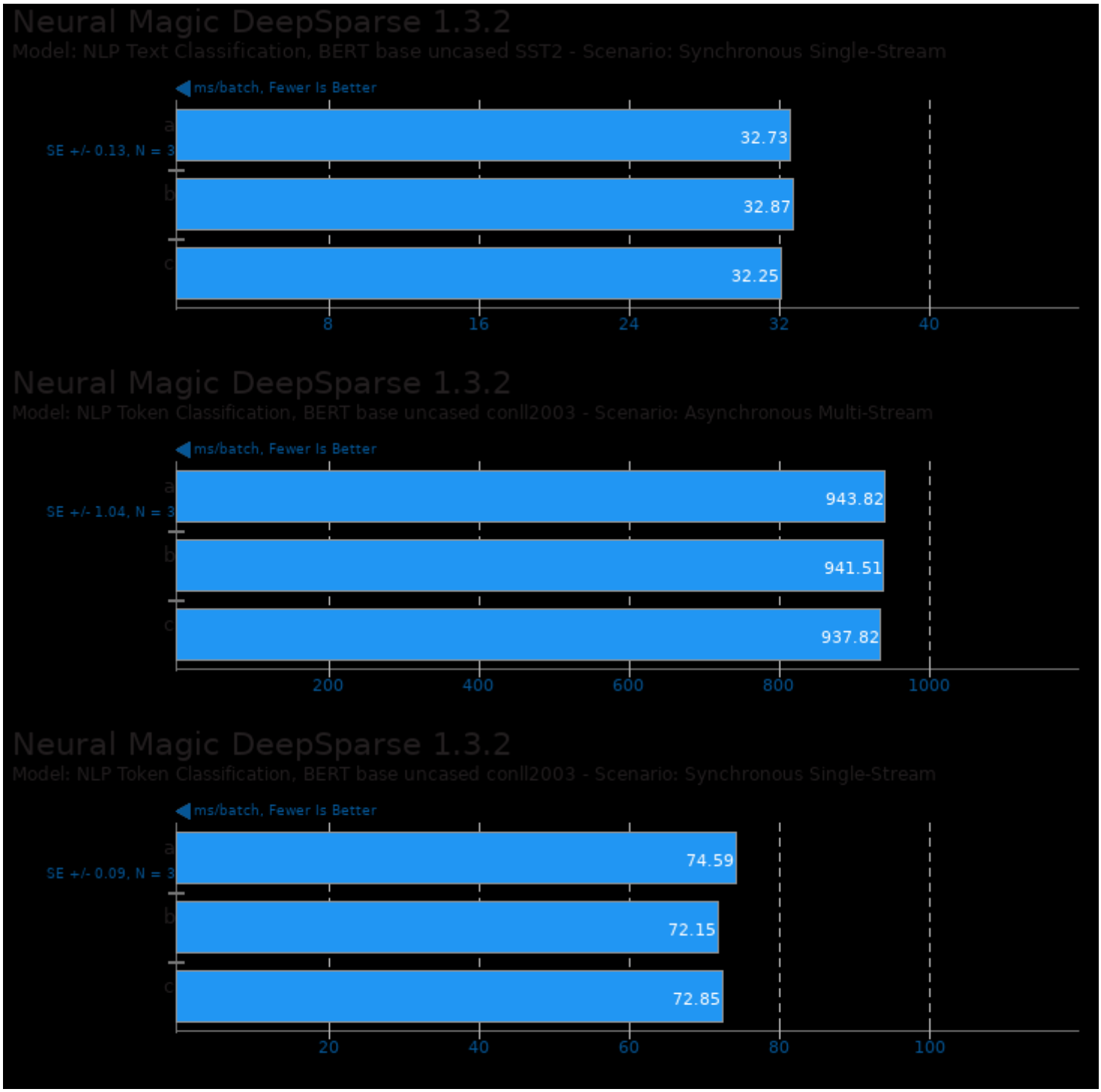

Neural Magic DeepSparse 1.3.2
Model: NLP Sentiment Analysis, 80% Pruned Quantized BERT Base Uncased - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.22, N = 3
a
21.08
b
20.11
c
20.13
5
10
15
20
25
Neural Magic DeepSparse 1.3.2
Model: NLP Question Answering, BERT base uncased SQuaD 12layer Pruned90 - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 1.72, N = 3
a
360.55
b
359.24
c
358.03
80
160
240
320
400
Neural Magic DeepSparse 1.3.2
Model: NLP Question Answering, BERT base uncased SQuaD 12layer Pruned90 - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.42, N = 3
a
48.01
b
47.35
c
47.08
11
22
33
44
55

deepsparse tr




Neural Magic DeepSparse 1.3.2
Model: CV Classification, ResNet-50 ImageNet - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.02, N = 3
a
10.56
b
10.51
c
10.67
3
6
9
12
15
Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, DistilBERT mnli - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 0.20, N = 3
a
114.46
b
112.57
c
112.89
30
60
90
120
150
Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, DistilBERT mnli - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.10, N = 3
a
17.28
b
16.63
c
16.80
4
8
12
16
20


Neural Magic DeepSparse 1.3.2
Model: CV Segmentation, 90% Pruned YOLACT Pruned - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 0.38, N = 3
a
b
c
728.34
725.26
724.52
160
320
480
640
800
Neural Magic DeepSparse 1.3.2
Model: CV Segmentation, 90% Pruned YOLACT Pruned - Scenario: Synchronous Single-Stream
ms/batch, Fewer Is Better
SE +/- 0.03, N = 3
a
b
c
54.62
54.55
54.45
12
24
36
48
60
Neural Magic DeepSparse 1.3.2
Model: NLP Text Classification, BERT base uncased SST2 - Scenario: Asynchronous Multi-Stream
ms/batch, Fewer Is Better
SE +/- 0.04, N = 3
a
b
c
224.91
222.26
222.60
50
100
150
200
250



This file was automatically generated via the Phoronix Test Suite benchmarking software on Friday, 29 March 2024 02:49.