Welvet QNN NPU lab · all · done · 182 rows · 4 fail(s) · 4 timeout(s)

mount = weights → NPU (create+finalize) · run = execute only · call = ABI wall

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
dense/float32	medium	ok	2.62ms	142.10ms	858.33µs	3.05× / +67.2%	0.0	1.0000	B=16 in=512 out=512
dense/float16	medium	ok	2.87ms	111.67ms	770.83µs	3.72× / +73.1%	—	1.0000	B=16 in=512 out=512
dense/bfloat16	medium	skip	—	—	—		—	—	—
dense/int8	medium	ok	2.00ms	119.71ms	686.35µs	2.92× / +65.8%	—	0.1483	B=16 in=512 out=512
dense/uint8	medium	ok	2.50ms	118.98ms	666.72µs	3.75× / +73.3%	—	0.1483	B=16 in=512 out=512
dense/int16	medium	ok	1.70ms	136.76ms	461.67µs	3.68× / +72.9%	—	0.1483	B=16 in=512 out=512
dense/uint16	medium	ok	1.93ms	118.29ms	716.72µs	2.70× / +62.9%	—	0.1483	B=16 in=512 out=512
softmax/float32	medium	ok	892.06µs	21.79ms	1.17ms	0.76× / -31.0%	—	1.0000	rows=32 cols=1024
softmax/float16	medium	ok	626.97µs	16.86ms	1.08ms	0.58× / -72.2%	—	1.0000	rows=32 cols=1024
softmax/bfloat16	medium	skip	—	—	—		—	—	—
softmax/int8	medium	ok	956.84µs	18.07ms	1.10ms	0.87× / -14.7%	—	-0.6947	rows=32 cols=1024
softmax/uint8	medium	ok	668.97µs	12.84ms	974.89µs	0.69× / -45.7%	—	1.0000	rows=32 cols=1024
softmax/int16	medium	ok	670.11µs	18.83ms	1.05ms	0.64× / -56.0%	—	1.0000	rows=32 cols=1024
softmax/uint16	medium	ok	923.26µs	15.18ms	1.09ms	0.85× / -18.3%	—	1.0000	rows=32 cols=1024
rmsnorm/float32	medium	ok	2.53ms	26.23ms	1.87ms	1.35× / +26.2%	0.0	1.0000	B=8 S=64 D=256
rmsnorm/float16	medium	ok	2.34ms	31.05ms	1.85ms	1.26× / +20.9%	0.0	1.0000	B=8 S=64 D=256
rmsnorm/bfloat16	medium	skip	—	—	—		—	—	—
rmsnorm/int8	medium	ok	2.34ms	17.29ms	1.61ms	1.45× / +31.0%	—	1.0000	B=8 S=64 D=256
rmsnorm/uint8	medium	ok	1.92ms	19.74ms	1.42ms	1.35× / +26.2%	—	1.0000	B=8 S=64 D=256
rmsnorm/int16	medium	ok	2.23ms	22.02ms	1.68ms	1.33× / +24.8%	—	1.0000	B=8 S=64 D=256
rmsnorm/uint16	medium	ok	2.35ms	23.29ms	1.55ms	1.51× / +33.8%	—	1.0000	B=8 S=64 D=256
layernorm/float32	medium	ok	2.93ms	24.30ms	1.38ms	2.13× / +53.0%	0.0	1.0000	B=8 S=64 D=256
layernorm/float16	medium	ok	2.73ms	18.82ms	1.20ms	2.28× / +56.1%	0.0	1.0000	B=8 S=64 D=256
layernorm/bfloat16	medium	skip	—	—	—		—	—	—
layernorm/int8	medium	ok	2.23ms	19.79ms	1.29ms	1.73× / +42.3%	—	1.0000	B=8 S=64 D=256

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
layernorm/uint8	medium	ok	2.53ms	16.63ms	1.33ms	1.90× / +47.3%	0	1.0000	B=8 S=64 D=256
layernorm/int16	medium	ok	1.98ms	23.29ms	1.25ms	1.58× / +36.8%	0	1.0000	B=8 S=64 D=256
layernorm/uint16	medium	ok	2.26ms	18.57ms	1.27ms	1.77× / +43.6%	0	1.0000	B=8 S=64 D=256
cnn2/float32	medium	ok	2.14ms	28.63ms	1.62ms	1.32× / +24.0%	0	-0.0060	N=4 32x32 cin=16 cout=32 k=1x1
cnn2/float16	medium	ok	2.51ms	25.42ms	1.44ms	1.74× / +42.7%	0	0.0014	N=4 32x32 cin=16 cout=32 k=1x1
cnn2/bfloat16	medium	skip	0	0	0		0	0	0
cnn2/int8	medium	ok	3.44ms	26.85ms	1.33ms	2.58× / +61.3%	0	1.0000	N=4 32x32 cin=16 cout=32 k=1x1
cnn2/uint8	medium	ok	2.43ms	28.54ms	1.37ms	1.77× / +43.6%	0	1.0000	N=4 32x32 cin=16 cout=32 k=1x1
cnn2/int16	medium	ok	2.10ms	25.97ms	2.15ms	0.98× / -2.4%	0	0.0029	N=4 32x32 cin=16 cout=32 k=1x1
cnn2/uint16	medium	ok	2.14ms	21.83ms	1.45ms	1.47× / +32.0%	0	1.0000	N=4 32x32 cin=16 cout=32 k=1x1
cnn3/float32	medium	ok	133.33µs	25.37ms	962.08µs	0.14× / -621.6%	0	0.0032	N=2 8x8x8 cin=8 cout=16
cnn3/float16	medium	ok	137.03µs	25.08ms	1.01ms	0.14× / -635.3%	0	0.0031	N=2 8x8x8 cin=8 cout=16
cnn3/bfloat16	medium	skip	0	0	0		0	0	0
cnn3/int8	medium	ok	170.31µs	27.79ms	1.10ms	0.16× / -543.4%	0	1.0000	N=2 8x8x8 cin=8 cout=16
cnn3/uint8	medium	ok	158.13µs	27.71ms	1.03ms	0.15× / -550.0%	0	1.0000	N=2 8x8x8 cin=8 cout=16
cnn3/int16	medium	skip	0	0	0		0	0	0
cnn3/uint16	medium	skip	0	0	0		0	0	0
convt2/float32	medium	ok	958.02µs	26.41ms	1.34ms	0.72× / -39.5%	0	0.0242	N=4 32x32 cin=16 cout=16
convt2/float16	medium	ok	1.45ms	28.65ms	1.28ms	1.13× / +11.3%	0	0.0004	N=4 32x32 cin=16 cout=16
convt2/bfloat16	medium	skip	0	0	0		0	0	0
convt2/int8	medium	ok	2.13ms	28.45ms	1.19ms	1.78× / +43.8%	0	1.0000	N=4 32x32 cin=16 cout=16
convt2/uint8	medium	ok	509.69µs	25.58ms	981.56µs	0.52× / -92.6%	0	1.0000	N=4 32x32 cin=16 cout=16
convt2/int16	medium	ok	1.16ms	29.45ms	1.53ms	0.76× / -31.3%	0	1.0000	N=4 32x32 cin=16 cout=16
convt2/uint16	medium	ok	1.65ms	29.93ms	1.31ms	1.26× / +20.6%	0	1.0000	N=4 32x32 cin=16 cout=16
convt3/float32	medium	ok	0	23.99ms	1.56ms	0	0	0	N=1 8x8x8 c=8[illegible]8
convt3/float16	medium	ok	0	18.57ms	1.01ms	0	0	0	N=1 8x8x8 c=8[illegible]8
convt3/bfloat16	medium	skip	0	0	0		0	0	0
convt3/int8	medium	ok	0	19.41ms	1.01ms	0	0	0	N=1 8x8x8 c=8[illegible]8

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
convt3/uint8	medium	ok	[illegible]	21.73ms	1.03ms	[illegible]	[illegible]	[illegible]	N=1 8x8x8 c=8[illegible]8
convt3/int16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
convt3/uint16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
lstm/float32	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	T=16 B=8 in=128 hid=128
lstm/float16	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	T=16 B=8 in=128 hid=128
lstm/bfloat16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
lstm/int8	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
lstm/uint8	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
lstm/int16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
lstm/uint16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
embedding/float32	medium	ok	21.93µs	25.88ms	1.04ms	0.02× / -4652.5%	0.0	1.0000	vocab=4096 dim=256 n=256
embedding/float16	medium	ok	165.21µs	15.90ms	1.01ms	0.16× / -514.0%	0.1	1.0000	vocab=4096 dim=256 n=256
embedding/bfloat16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
embedding/int8	medium	ok	29.95µs	18.00ms	1.03ms	0.03× / -3336.5%	[illegible]	1.0000	vocab=4096 dim=256 n=256
embedding/uint8	medium	ok	31.41µs	16.96ms	1.02ms	0.03× / -3148.1%	[illegible]	1.0000	vocab=4096 dim=256 n=256
embedding/int16	medium	ok	34.17µs	17.13ms	1.14ms	0.03× / -3239.0%	[illegible]	1.0000	vocab=4096 dim=256 n=256
embedding/uint16	medium	ok	28.65µs	12.77ms	1.08ms	0.03× / -3656.5%	[illegible]	1.0000	vocab=4096 dim=256 n=256
residual/float32	medium	ok	105.26µs	1934.71ms	3.34ms	0.03× / -3075.5%	0.0	1.0000	n=65536
residual/float16	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	single
residual/bfloat16	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	single
residual/int8	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	single
residual/uint8	medium	fail	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	single
residual/int16	medium	ok	191.30µs	19.24ms	1.24ms	0.15× / -547.5%	[illegible]	1.0000	n=65536
residual/uint16	medium	ok	85.05µs	12.11ms	1.10ms	0.08× / -1196.1%	[illegible]	1.0000	n=65536
swiglu/float32	medium	ok	2.50ms	26.15ms	1.17ms	2.13× / +53.0%	[illegible]	-0.0021	B=8 in=256 hidden=768
swiglu/float16	medium	ok	2.50ms	16.57ms	997.97µs	2.50× / +60.0%	[illegible]	-0.0020	B=8 in=256 hidden=768
swiglu/bfloat16	medium	skip	[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	[illegible]
swiglu/int8	medium	ok	2.48ms	19.21ms	1.07ms	2.31× / +56.8%	[illegible]	1.0000	B=8 in=256 hidden=768

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
swiglu/uint8	medium	ok	3.45ms	21.68ms	1.06ms	3.26× / +69.3%	✗	1.0000	B=8 in=256 hidden=768
swiglu/int16	medium	ok	2.53ms	20.92ms	1.07ms	2.36× / +57.6%	✗	1.0000	B=8 in=256 hidden=768
swiglu/uint16	medium	ok	2.71ms	21.56ms	1.09ms	2.48× / +59.8%	✗	1.0000	B=8 in=256 hidden=768
mha/float32	medium	ok	25.96ms	24.77ms	1.18ms	22.07× / +95.5%	✗	0.4872	B=2 S=64 D=256 H=8
mha/float16	medium	ok	30.99ms	17.32ms	1.13ms	27.32× / +96.3%	✗	0.4872	B=2 S=64 D=256 H=8
mha/bfloat16	medium	skip	✗	✗	✗		✗	✗	✗
mha/int8	medium	ok	32.17ms	18.80ms	1.22ms	26.45× / +96.2%	✗	1.0000	B=2 S=64 D=256 H=8
mha/uint8	medium	ok	34.10ms	14.05ms	906.04µs	37.64× / +97.3%	✗	1.0000	B=2 S=64 D=256 H=8
mha/int16	medium	ok	27.98ms	22.60ms	1.25ms	22.31× / +95.5%	✗	-0.0656	B=2 S=64 D=256 H=8
mha/uint16	medium	ok	31.37ms	19.76ms	1.18ms	26.56× / +96.2%	✗	1.0000	B=2 S=64 D=256 H=8
stack/dense/float32	medium/stack	ok	435.84µs	153.18µs	100.96ms	0.00× / -23065.0%	✗	1.0000	Dense×5 B=4 dim=96
stack/dense/float16	medium/stack	ok	1.28ms	12.26ms	2.43ms	0.52× / -90.8%	✗	1.0000	dense×5 B=4 in=192 out=192
stack/dense/bfloat16	medium/stack	skip	✗	✗	✗		✗	✗	dense×5 ✗
stack/dense/int8	medium/stack	ok	899.48µs	12.45ms	4.07ms	0.22× / -352.7%	✗	0.2457	dense×5 B=4 in=192 out=192
stack/dense/uint8	medium/stack	ok	927.53µs	12.85ms	2.15ms	0.43× / -131.8%	✗	0.2457	dense×5 B=4 in=192 out=192
stack/dense/int16	medium/stack	ok	968.83µs	18.67ms	1.96ms	0.49× / -102.2%	✗	0.2457	dense×5 B=4 in=192 out=192
stack/dense/uint16	medium/stack	ok	1.56ms	13.65ms	3.74ms	0.42× / -139.6%	✗	0.2457	dense×5 B=4 in=192 out=192
stack/softmax/float32	medium/stack	ok	107.46µs	14.74ms	1.10ms	0.10× / -924.6%	✗	1.0000	Softmax×5 B=4 dim=96
stack/softmax/float16	medium/stack	ok	488.83µs	16.85ms	5.94ms	0.08× / -1114.2%	✗	1.0000	softmax×5 rows=8 cols=256
stack/softmax/bfloat16	medium/stack	skip	✗	✗	✗		✗	✗	softmax×5 ✗
stack/softmax/int8	medium/stack	ok	642.88µs	16.61ms	5.49ms	0.12× / -754.3%	✗	-0.6901	softmax×5 rows=8 cols=256
stack/softmax/uint8	medium/stack	ok	597.57µs	13.16ms	6.48ms	0.09× / -984.1%	✗	1.0000	softmax×5 rows=8 cols=256
stack/softmax/int16	medium/stack	ok	554.26µs	15.10ms	5.75ms	0.10× / -937.6%	✗	1.0000	softmax×5 rows=8 cols=256
stack/softmax/uint16	medium/stack	ok	207.42µs	17.19ms	5.83ms	0.04× / -2708.3%	✗	1.0000	softmax×5 rows=8 cols=256
stack/rmsnorm/float32	medium/stack	ok	43.75µs	35.52µs	98.95ms	0.00× / -226101.8%	✗	1.0000	RMSNorm×5 B=4 dim=96
stack/rmsnorm/float16	medium/stack	ok	605.50µs	17.00ms	5.73ms	0.11× / -846.1%	✗	1.0000	rmsnorm×5 B=2 S=16 D=96
stack/rmsnorm/bfloat16	medium/stack	skip	✗	✗	✗		✗	✗	rmsnorm×5 ✗
stack/rmsnorm/int8	medium/stack	ok	296.82µs	11.01ms	4.30ms	0.07× / -1349.4%	✗	1.0000	rmsnorm×5 B=2 S=16 D=96

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
stack/rmsnorm/uint8	medium/stack	ok	657.92µs	15.82ms	5.60ms	0.12× / -751.8%	□	1.0000	rmsnorm×5 B=2 S=16 D=96
stack/rmsnorm/int16	medium/stack	ok	681.97µs	18.30ms	5.75ms	0.12× / -743.1%	□	1.0000	rmsnorm×5 B=2 S=16 D=96
stack/rmsnorm/uint16	medium/stack	ok	614.47µs	17.26ms	5.68ms	0.11× / -823.8%	□	1.0000	rmsnorm×5 B=2 S=16 D=96
stack/layernorm/float32	medium/stack	ok	40.91µs	124.58µs	87.27ms	0.00× / -213243.7%	□	1.0000	LayerNorm×5 B=4 dim=96
stack/layernorm/float16	medium/stack	ok	673.55µs	16.12ms	5.15ms	0.13× / -665.3%	0.0	1.0000	layernorm×5 B=2 S=16 D=96
stack/layernorm/bfloat16	medium/stack	skip	□	□	□		□	□	layernorm×5 □
stack/layernorm/int8	medium/stack	ok	272.93µs	14.58ms	5.40ms	0.05× / -1878.5%	□	1.0000	layernorm×5 B=2 S=16 D=96
stack/layernorm/uint8	medium/stack	ok	262.44µs	15.42ms	5.04ms	0.05× / -1821.5%	□	1.0000	layernorm×5 B=2 S=16 D=96
stack/layernorm/int16	medium/stack	ok	278.11µs	20.19ms	5.72ms	0.05× / -1955.2%	□	1.0000	layernorm×5 B=2 S=16 D=96
stack/layernorm/uint16	medium/stack	ok	643.77µs	17.01ms	5.63ms	0.11× / -775.1%	□	1.0000	layernorm×5 B=2 S=16 D=96
stack/cnn2/float32	medium/stack	ok	34.27µs	14.21ms	3.42ms	0.01× / -9889.3%	□	0.0206	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn2/float16	medium/stack	ok	67.34µs	15.70ms	3.27ms	0.02× / -4750.8%	□	0.0208	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn2/bfloat16	medium/stack	skip	□	□	□		□	□	cnn2×3 □
stack/cnn2/int8	medium/stack	ok	58.54µs	17.49ms	3.27ms	0.02× / -5483.9%	□	1.0000	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn2/uint8	medium/stack	ok	67.03µs	17.33ms	3.24ms	0.02× / -4732.4%	□	1.0000	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn2/int16	medium/stack	ok	29.32µs	13.60ms	3.21ms	0.01× / -10859.8%	□	1.0000	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn2/uint16	medium/stack	ok	67.24µs	17.57ms	2.96ms	0.02× / -4294.9%	□	1.0000	cnn2×3 N=1 12x12 cin=8 cout=8 k=1x1
stack/cnn3/float32	medium/stack	ok	592.34µs	26.45ms	3.25ms	0.18× / -449.1%	□	0.2752	cnn3×3 N=1 4x4x4 cin=4 cout=4
stack/cnn3/float16	medium/stack	ok	9.063µs	26.73ms	2.84ms	0.00× / -31188.0%	□	0.2751	cnn3×3 N=1 4x4x4 cin=4 cout=4
stack/cnn3/bfloat16	medium/stack	skip	□	□	□		□	□	cnn3×3 □
stack/cnn3/int8	medium/stack	ok	6.405µs	15.38ms	2.58ms	0.00× / -40171.3%	□	1.0000	cnn3×3 N=1 4x4x4 cin=4 cout=4
stack/cnn3/uint8	medium/stack	ok	9.846µs	22.43ms	3.13ms	0.00× / -31737.2%	□	1.0000	cnn3×3 N=1 4x4x4 cin=4 cout=4
stack/cnn3/int16	medium/stack	skip	□	□	□		□	□	cnn3×3 □
stack/cnn3/uint16	medium/stack	skip	□	□	□		□	□	cnn3×3 □
stack/convt2/float32	medium/stack	ok	530.94µs	18.34ms	3.10ms	0.17× / -483.5%	□	-0.0699	convt2×3 N=1 12x12 cin=8 cout=8
stack/convt2/float16	medium/stack	ok	646.09µs	18.52ms	3.24ms	0.20× / -401.5%	□	0.0056	convt2×3 N=1 12x12 cin=8 cout=8
stack/convt2/bfloat16	medium/stack	skip	□	□	□		□	□	convt2×3 □
stack/convt2/int8	medium/stack	ok	270.16µs	11.91ms	2.60ms	0.10× / -862.3%	□	1.0000	convt2×3 N=1 12x12 cin=8 cout=8

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
stack/convt2/uint8	medium/stack	ok	510.78µs	18.75ms	2.91ms	0.18× / -470.6%	□	1.0000	convt2×3 N=1 12x12 cin=8 cout=8
stack/convt2/int16	medium/stack	ok	70.47µs	17.25ms	3.10ms	0.02× / -4293.5%	□	1.0000	convt2×3 N=1 12x12 cin=8 cout=8
stack/convt2/uint16	medium/stack	ok	448.44µs	17.43ms	3.03ms	0.15× / -575.9%	□	1.0000	convt2×3 N=1 12x12 cin=8 cout=8
stack/convt3/float32	medium/stack	ok	□	24.68ms	3.01ms	□	□	□	convt3×3 N=1 4x4x4 c=4□4
stack/convt3/float16	medium/stack	ok	□	25.40ms	3.04ms	□	□	□	convt3×3 N=1 4x4x4 c=4□4
stack/convt3/bfloat16	medium/stack	skip	□	□	□		□	□	convt3×3 □
stack/convt3/int8	medium/stack	ok	□	16.07ms	2.85ms	□	□	□	convt3×3 N=1 4x4x4 c=4□4
stack/convt3/uint8	medium/stack	ok	□	20.56ms	3.27ms	□	□	□	convt3×3 N=1 4x4x4 c=4□4
stack/convt3/int16	medium/stack	skip	□	□	□		□	□	convt3×3 □
stack/convt3/uint16	medium/stack	skip	□	□	□		□	□	convt3×3 □
stack/lstm/float32	medium/stack	fail	□	□	□		□	□	lstm×3 T=4 B=2 in=48 hid=48
stack/lstm/float16	medium/stack	fail	□	□	□		□	□	lstm×3 T=4 B=2 in=48 hid=48
stack/lstm/bfloat16	medium/stack	skip	□	□	□		□	□	lstm×3 □
stack/lstm/int8	medium/stack	skip	□	□	□		□	□	lstm×3 □
stack/lstm/uint8	medium/stack	skip	□	□	□		□	□	lstm×3 □
stack/lstm/int16	medium/stack	skip	□	□	□		□	□	lstm×3 □
stack/lstm/uint16	medium/stack	skip	□	□	□		□	□	lstm×3 □
stack/embedding/float32	medium/stack	ok	4.531µs	12.28ms	6.03ms	0.00× / -132878.7%	□	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/embedding/float16	medium/stack	ok	12.50µs	15.02ms	5.54ms	0.00× / -44197.9%	0.1	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/embedding/bfloat16	medium/stack	skip	□	□	□		□	□	embedding×5 □
stack/embedding/int8	medium/stack	ok	9.896µs	16.61ms	5.73ms	0.00× / -57759.5%	□	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/embedding/uint8	medium/stack	ok	4.062µs	11.66ms	5.55ms	0.00× / -136410.5%	□	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/embedding/int16	medium/stack	ok	9.896µs	13.93ms	5.55ms	0.00× / -56022.7%	□	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/embedding/uint16	medium/stack	ok	12.50µs	17.02ms	5.78ms	0.00× / -46177.1%	□	1.0000	embedding×5 vocab=512 dim=64 n=64
stack/residual/float32	medium/stack	ok	27.86µs	161.62ms	15.57ms	0.00× / -55767.4%	□	1.0000	residual×5 n=8192
stack/residual/float16	medium/stack	ok	27.24µs	153.83ms	11.65ms	0.00× / -42653.7%	0.0	1.0000	residual×5 n=8192
stack/residual/bfloat16	medium/stack	skip	□	□	□		□	□	residual×5 □

Layer	Size	Status	Go (CPU)	NPU mount	NPU run	Speedup	Exact%	Cos	Geom
stack/residual/int8	medium/stack	ok	57.34µs	11.21ms	11.71ms	0.00× / -20315.9%	☒	1.0000	residual×5 n=8192
stack/residual/uint8	medium/stack	ok	88.18µs	14.65ms	4.78ms	0.02× / -5324.7%	☒	1.0000	residual×5 n=8192
stack/residual/int16	medium/stack	ok	88.75µs	20.33ms	6.03ms	0.01× / -6699.0%	☒	1.0000	residual×5 n=8192
stack/residual/uint16	medium/stack	ok	41.93µs	13.92ms	5.52ms	0.01× / -13057.2%	☒	1.0000	residual×5 n=8192
stack/swiglu/float32	medium/stack	ok	1.40ms	21.02ms	3.49ms	0.40× / -148.7%	☒	-0.1079	swiglu×3 B=2 in=96 hidden=256
stack/swiglu/float16	medium/stack	ok	1.03ms	17.58ms	3.25ms	0.32× / -216.5%	☒	-0.1058	swiglu×3 B=2 in=96 hidden=256
stack/swiglu/bfloat16	medium/stack	skip	☒	☒	☒		☒	☒	swiglu×3 ☒
stack/swiglu/int8	medium/stack	ok	1.37ms	19.96ms	3.48ms	0.39× / -153.5%	☒	1.0000	swiglu×3 B=2 in=96 hidden=256
stack/swiglu/uint8	medium/stack	ok	1.51ms	13.73ms	2.75ms	0.55× / -82.0%	☒	1.0000	swiglu×3 B=2 in=96 hidden=256
stack/swiglu/int16	medium/stack	ok	1.47ms	20.00ms	3.37ms	0.44× / -128.7%	☒	1.0000	swiglu×3 B=2 in=96 hidden=256
stack/swiglu/uint16	medium/stack	ok	886.65µs	17.01ms	3.15ms	0.28× / -254.9%	☒	1.0000	swiglu×3 B=2 in=96 hidden=256
stack/mha/float32	medium/stack	ok	2.85ms	20.20ms	3.77ms	0.76× / -32.4%	☒	0.4871	mha×3 B=1 S=16 D=64 H=4
stack/mha/float16	medium/stack	ok	2.67ms	19.18ms	3.46ms	0.77× / -29.5%	☒	0.4871	mha×3 B=1 S=16 D=64 H=4
stack/mha/bfloat16	medium/stack	skip	☒	☒	☒		☒	☒	mha×3 ☒
stack/mha/int8	medium/stack	ok	2.66ms	19.77ms	3.49ms	0.76× / -31.6%	☒	1.0000	mha×3 B=1 S=16 D=64 H=4
stack/mha/uint8	medium/stack	ok	2.52ms	18.81ms	3.44ms	0.73× / -36.7%	☒	1.0000	mha×3 B=1 S=16 D=64 H=4
stack/mha/int16	medium/stack	ok	2.59ms	20.75ms	3.44ms	0.75× / -32.6%	☒	-0.0004	mha×3 B=1 S=16 D=64 H=4
stack/mha/uint16	medium/stack	ok	5.11ms	15.07ms	3.12ms	1.64× / +38.9%	☒	1.0000	mha×3 B=1 S=16 D=64 H=4