

Appendix B Model Kernel Fp8 (Page 1/4)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
RTX4090	Llama3.2-1B	64	6.64e+01	0.80	0.78	0.71	0.63	0.61	0.50	0.65	0.52
		128	8.19e+01	1.03	0.91	0.84	0.74	0.69	0.58	0.70	0.64
		256	1.27e+02	1.45	1.14	1.02	0.97	0.45	0.51	0.52	0.42
		512	2.17e+02	1.35	1.27	1.19	1.07	0.46	0.49	0.50	0.44
		1024	4.07e+02	1.69	1.41	1.30	1.19	0.56	0.59	0.48	0.45
		2048	7.98e+02	1.77	1.46	1.35	1.20	0.55	0.52	0.83	0.41
		4096	1.59e+03	1.97	1.45	1.35	1.23	0.54	0.95	1.01	0.81
		8192	3.21e+03	1.98	1.49	1.36	1.24	0.93	0.95	1.04	0.91
		16384	6.46e+03	2.00	1.49	1.36	1.24	1.19	1.18	1.17	1.03
	BitNet-2B	64	6.61e+01	0.84	0.66	0.68	0.67	0.49	0.61	0.59	0.54
		128	9.20e+01	1.14	0.81	0.90	0.88	0.73	0.80	0.78	0.71
		256	1.51e+02	1.55	1.14	1.12	1.12	0.95	1.00	1.04	0.90
		512	2.62e+02	1.62	1.24	1.14	1.08	1.03	1.00	1.00	0.88
		1024	4.90e+02	1.68	1.26	1.16	1.10	1.05	1.02	1.01	0.89
		2048	8.93e+02	1.77	1.33	1.21	1.15	1.09	1.04	1.05	0.92
H100	Llama3.2-3B	4096	1.80e+03	1.91	1.42	1.30	1.22	1.16	1.10	1.12	0.98
		8192	3.58e+03	1.96	1.45	1.34	1.25	1.19	1.14	1.14	1.00
		16384	7.26e+03	2.00	1.48	1.35	1.27	1.20	1.16	1.15	1.02
		64	7.80e+01	0.85	0.72	0.70	0.60	0.62	0.57	0.27	0.58
		128	1.17e+02	1.34	0.94	0.88	0.94	0.62	0.48	0.49	0.40
		256	2.03e+02	1.53	1.35	1.24	1.17	0.43	0.48	0.29	0.34
		512	3.58e+02	1.29	1.33	1.15	0.85	0.34	0.31	0.51	0.40
		1024	6.58e+02	1.46	1.34	1.20	1.10	0.28	0.41	0.46	0.63
		2048	1.32e+03	2.01	1.52	1.36	1.26	0.60	1.07	0.63	0.58
		4096	2.64e+03	2.03	1.54	1.39	1.27	0.83	1.17	0.86	0.59
	Qwen2.5-7B	8192	5.31e+03	2.04	1.54	1.38	1.28	1.23	1.15	1.18	1.03
		16384	1.07e+04	2.05	1.56	1.39	1.29	1.24	1.19	1.20	1.06
		64	2.33e+02	1.17	0.88	0.80	0.71	0.24	0.44	0.34	0.45
		128	2.83e+02	1.37	0.93	0.89	0.88	0.29	0.28	0.19	0.34
	Qwen2.5-14B	256	4.69e+02	1.14	1.36	1.23	1.16	0.52	0.51	0.17	0.47
		512	8.31e+02	1.80	1.43	1.29	1.20	0.46	0.43	0.38	0.48
		1024	1.62e+03	1.93	1.46	1.32	1.24	0.55	0.63	0.42	0.43
		2048	3.07e+03	1.91	1.43	1.29	1.20	0.81	1.11	0.77	0.98
		4096	6.12e+03	2.00	1.49	1.36	1.27	1.16	1.07	1.08	1.00
		8192	1.24e+04	2.04	1.52	1.38	1.29	1.22	1.22	1.20	1.06
		16384	2.48e+04	2.05	1.52	1.37	1.28	1.21	1.21	1.20	1.04
		64	2.59e+02	1.36	1.04	0.92	0.87	0.32	0.13	0.29	0.31
		128	3.22e+02	1.59	0.80	1.07	0.99	0.21	0.31	0.18	0.38
		256	5.36e+02	1.34	1.31	1.09	1.01	0.21	0.29	0.30	0.27
	Llama3.2-1B	512	1.00e+03	1.76	1.22	1.10	1.06	0.30	0.34	0.32	0.30
		1024	1.82e+03	1.87	1.39	1.25	1.13	0.70	0.71	0.66	0.79
		2048	3.57e+03	1.99	1.48	1.32	1.19	1.13	1.10	0.79	1.02
		4096	7.14e+03	2.03	1.52	1.36	1.19	1.18	1.11	1.16	1.02
		8192	1.45e+04	2.06	1.54	1.38	1.30	1.17	1.19	1.19	1.05
		16384	2.95e+04	2.10	1.55	1.40	1.31	1.25	1.21	1.22	1.07
H100	BitNet-2B	64	3.80e+01	0.95	0.81	0.78	0.72	0.70	0.71	0.73	0.60
		128	4.27e+01	0.97	0.81	0.80	0.72	0.69	0.70	0.71	0.60
		256	5.12e+01	0.91	0.73	0.69	0.64	0.60	0.60	0.61	0.55
		512	8.43e+01	1.04	0.79	0.79	0.67	0.65	0.68	0.69	0.64
		1024	1.51e+02	1.17	0.95	0.86	0.83	0.75	0.77	0.78	0.71
		2048	2.88e+02	1.24	0.98	0.95	0.83	0.80	0.78	0.80	0.72
		4096	5.83e+02	1.37	1.06	0.98	0.86	0.87	0.87	0.84	0.78
		8192	1.13e+03	1.42	1.07	0.94	0.85	0.83	0.85	0.82	0.72
		16384	2.53e+03	1.54	1.14	1.02	0.96	0.91	0.87	0.90	0.81
		64	4.52e+01	1.18	0.95	0.99	0.95	0.91	0.87	0.89	0.69
	Llama3.2-3B	128	4.62e+01	1.10	0.87	0.91	0.88	0.81	0.77	0.76	0.65
		256	5.63e+01	1.04	0.85	0.81	0.76	0.74	0.67	0.69	0.62
		512	8.86e+01	1.10	0.87	0.84	0.77	0.74	0.70	0.71	0.65
		1024	1.67e+02	1.25	0.93	0.90	0.85	0.81	0.80	0.78	0.69
		2048	3.19e+02	1.30	1.00	0.95	0.85	0.86	0.79	0.82	0.72
		4096	6.87e+02	1.48	1.12	1.05	0.99	0.93	0.89	0.90	0.82
	Llama3.2-1B	8192	1.36e+03	1.50	1.11	1.04	0.98	0.94	0.90	0.88	0.78
		16384	2.88e+03	1.57	1.11	1.05	0.99	0.92	0.88	0.92	0.80
	BitNet-2B	64	6.69e+01	1.46	1.11	0.97	0.92	0.88	0.85	0.88	0.82
		128	6.64e+01	1.27	0.96	0.88	0.81	0.80	0.75	0.79	0.71

≥1.5x ≥1.0x

Appendix B Model Kernel Fp8 (Page 2/4)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		256	7.85e+01	1.07	0.87	0.77	0.72	0.71	0.67	0.67	0.61
		512	1.29e+02	1.20	0.93	0.86	0.81	0.77	0.70	0.75	0.66
		1024	2.47e+02	1.39	1.07	0.97	0.91	0.90	0.85	0.84	0.75
		2048	4.78e+02	1.49	1.16	1.04	0.96	0.92	0.85	0.89	0.77
		4096	1.09e+03	1.67	1.34	1.19	1.08	1.03	0.97	1.02	0.87
		8192	1.95e+03	1.53	1.22	1.02	0.95	0.91	0.82	0.89	0.76
		16384	4.66e+03	1.75	1.35	1.20	1.12	1.02	0.94	1.03	0.89
Qwen2.5-7B	Qwen2.5-7B	64	1.53e+02	1.53	1.16	1.08	1.00	0.97	0.97	0.95	0.85
		128	1.46e+02	1.29	1.00	0.94	0.89	0.81	0.83	0.81	0.74
		256	1.80e+02	1.26	0.94	0.90	0.85	0.78	0.78	0.79	0.72
		512	3.05e+02	1.32	1.08	0.96	0.90	0.84	0.84	0.84	0.77
		1024	5.61e+02	1.54	1.14	1.08	0.97	0.87	0.90	0.92	0.81
		2048	1.14e+03	1.60	1.23	1.08	1.03	0.84	0.90	0.93	0.82
		4096	2.55e+03	1.74	1.25	1.13	0.99	0.96	0.93	0.94	0.85
		8192	4.96e+03	1.60	1.19	1.08	0.93	0.89	0.92	0.92	0.82
		16384	1.10e+04	1.74	1.30	1.19	1.02	1.00	1.01	1.01	0.90
		64	1.48e+02	1.38	1.06	0.99	0.94	0.87	0.83	0.85	0.72
Qwen2.5-14B	Qwen2.5-14B	128	1.56e+02	1.32	1.00	0.94	0.89	0.80	0.77	0.81	0.70
		256	1.92e+02	1.31	0.98	0.75	0.85	0.80	0.75	0.79	0.70
		512	3.69e+02	1.58	1.18	1.05	1.00	0.94	0.90	0.94	0.82
		1024	7.09e+02	1.50	1.22	1.12	1.05	0.98	0.90	0.99	0.84
		2048	1.48e+03	1.84	1.35	1.24	1.16	1.02	0.94	1.06	0.89
		4096	2.99e+03	1.72	1.33	1.14	1.11	1.01	0.94	0.99	0.88
		8192	6.30e+03	1.73	1.34	1.08	1.01	1.01	0.95	1.03	0.88
		16384	1.30e+04	1.67	1.32	1.07	1.00	1.02	1.00	1.02	0.90
		64	3.64e+01	1.17	1.01	1.01	0.93	0.88	0.93	0.95	0.93
		128	3.78e+01	1.22	0.99	1.02	0.96	0.92	0.92	0.92	0.92
B200	Llama3.2-1B	256	4.32e+01	1.28	1.05	1.05	1.02	0.95	0.99	1.00	0.99
		512	4.96e+01	1.23	1.08	1.05	0.94	0.86	0.95	1.01	0.93
		1024	6.33e+01	1.25	1.05	1.03	0.90	0.88	0.88	0.93	0.89
		2048	9.71e+01	1.35	1.05	1.04	0.92	0.85	0.88	0.93	0.86
		4096	1.70e+02	1.43	1.05	1.03	0.87	0.80	0.84	0.92	0.82
		8192	3.42e+02	1.58	1.16	1.12	0.97	0.90	0.93	0.99	0.88
		16384	6.55e+02	1.61	1.15	1.14	0.95	0.89	0.91	0.99	0.86
		64	3.50e+01	1.10	0.94	0.94	0.94	0.94	0.87	0.94	0.91
		128	3.39e+01	1.04	0.91	0.87	0.86	0.86	0.82	0.86	0.85
		256	3.74e+01	1.03	0.94	0.91	0.90	0.89	0.83	0.87	0.81
BitNet-2B	BitNet-2B	512	4.79e+01	1.13	0.89	0.94	0.91	0.86	0.83	0.86	0.83
		1024	6.52e+01	1.18	0.91	0.94	0.93	0.87	0.84	0.86	0.80
		2048	1.05e+02	1.31	0.93	0.98	0.92	0.85	0.77	0.85	0.78
		4096	1.90e+02	1.41	1.05	1.03	0.98	0.90	0.85	0.90	0.79
		8192	3.65e+02	1.54	1.08	1.08	1.01	0.94	0.88	0.93	0.83
		16384	7.37e+02	1.62	1.13	1.13	1.05	0.96	0.92	0.95	0.86
		64	4.16e+01	1.20	1.12	1.08	1.05	1.01	0.86	1.02	1.00
		128	4.03e+01	1.14	1.03	1.03	0.95	0.97	0.83	0.97	0.88
		256	4.45e+01	1.13	1.04	0.98	0.93	0.90	0.83	0.93	0.84
		512	5.71e+01	1.20	1.03	0.96	0.88	0.89	0.79	0.89	0.84
Qwen2.5-7B	Qwen2.5-7B	1024	8.24e+01	1.22	1.04	0.95	0.87	0.88	0.79	0.85	0.79
		2048	1.43e+02	1.37	1.12	1.02	0.89	0.89	0.79	0.89	0.79
		4096	2.75e+02	1.56	1.24	1.12	0.98	0.99	0.87	0.97	0.86
		8192	5.41e+02	1.64	1.26	1.13	0.99	1.00	0.88	0.99	0.87
		16384	1.11e+03	1.72	1.32	1.20	1.06	1.04	0.92	1.03	0.92
		64	6.08e+01	1.34	1.05	1.01	0.98	0.86	0.89	0.91	0.83
		128	6.30e+01	1.33	1.02	0.99	0.96	0.84	0.87	0.87	0.81
		256	7.33e+01	1.41	1.08	1.01	0.96	0.84	0.89	0.91	0.82
		512	1.01e+02	1.37	1.03	1.01	0.98	0.80	0.87	0.89	0.80
		1024	1.71e+02	1.49	1.12	1.07	1.01	0.86	0.94	0.93	0.84
Qwen2.5-14B	Qwen2.5-14B	2048	3.33e+02	1.58	1.13	1.12	1.05	0.84	0.93	0.98	0.86
		4096	6.18e+02	1.60	1.13	1.09	1.04	0.86	0.93	0.95	0.86
		8192	1.28e+03	1.75	1.22	1.15	1.09	0.93	1.01	0.99	0.89
		16384	2.59e+03	1.79	1.26	1.18	1.13	0.93	1.02	1.07	0.90
		64	6.84e+01	1.55	1.10	1.07	1.01	0.92	0.85	0.95	0.86
		128	7.23e+01	1.51	1.13	1.09	1.06	0.90	0.86	0.96	0.87
		256	7.80e+01	1.31	1.00	0.98	0.93	0.81	0.77	0.87	0.76
		512	1.15e+02	1.34	1.06	1.02	0.96	0.84	0.79	0.89	0.79

Appendix B Model Kernel Fp8 (Page 3/4)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
RTX5080	Llama3.2-1B	1024	2.01e+02	1.47	1.07	1.03	0.97	0.82	0.77	0.89	0.80
		2048	3.73e+02	1.59	1.16	1.09	1.03	0.90	0.85	0.94	0.83
		4096	7.34e+02	1.72	1.23	1.16	1.07	0.94	0.86	0.98	0.86
		8192	1.53e+03	1.85	1.32	1.25	1.16	1.01	0.94	1.06	0.92
		16384	3.12e+03	1.84	1.28	1.19	1.14	1.01	0.94	1.05	0.92
		64	5.57e+01	0.89	0.76	0.75	0.67	0.66	0.67	0.67	0.62
		128	9.48e+01	1.22	1.00	0.95	0.88	0.85	0.85	0.84	0.76
		256	1.62e+02	1.50	1.17	1.06	0.95	0.90	0.93	0.92	0.85
		512	2.97e+02	1.56	1.20	1.11	1.01	0.94	0.96	0.98	0.88
		1024	5.64e+02	1.64	1.24	1.15	1.06	1.00	1.00	1.00	0.89
BitNet-2B	BitNet-2B	2048	1.09e+03	1.69	1.29	1.15	1.08	1.03	1.03	1.00	0.88
		4096	2.14e+03	1.70	1.29	1.17	1.09	1.03	1.03	1.00	0.89
		8192	4.22e+03	1.70	1.28	1.15	1.09	1.03	1.02	1.00	0.88
		16384	8.40e+03	1.69	1.29	1.15	1.08	1.02	1.01	1.00	0.87
		64	6.17e+01	1.02	0.84	0.84	0.82	0.77	0.75	0.77	0.72
		128	9.66e+01	1.28	0.98	0.96	0.94	0.87	0.84	0.86	0.79
		256	1.72e+02	1.46	1.13	1.04	1.00	0.95	0.90	0.90	0.82
		512	3.20e+02	1.60	1.15	1.12	1.07	1.00	0.96	0.96	0.86
		1024	6.32e+02	1.70	1.27	1.17	1.11	1.05	1.01	1.00	0.89
		2048	1.23e+03	1.71	1.29	1.17	1.10	1.06	1.02	1.00	0.89
Llama3.2-3B	Llama3.2-3B	4096	2.43e+03	1.73	1.28	1.17	1.11	1.06	1.01	1.00	0.89
		8192	4.80e+03	1.72	1.27	1.15	1.09	1.05	1.00	0.99	0.87
		16384	9.53e+03	1.70	1.25	1.15	1.08	1.03	0.98	0.98	0.87
		64	8.99e+01	1.20	0.99	0.94	0.86	0.85	0.76	0.86	0.74
		128	1.43e+02	1.44	1.14	1.03	0.96	0.95	0.88	0.95	0.83
		256	2.53e+02	1.50	1.19	1.08	1.00	0.97	0.92	0.95	0.85
		512	4.68e+02	1.61	1.26	1.12	1.02	1.02	0.95	0.98	0.85
		1024	9.10e+02	1.73	1.32	1.17	1.07	1.06	1.01	1.00	0.88
		2048	1.77e+03	1.71	1.30	1.17	1.08	1.06	1.03	1.00	0.89
		4096	3.50e+03	1.73	1.32	1.16	1.08	1.07	1.03	1.00	0.88
Qwen2.5-7B	Qwen2.5-7B	8192	6.93e+03	1.72	1.30	1.16	1.07	1.05	1.03	1.00	0.88
		16384	1.38e+04	1.71	1.30	1.16	1.07	1.05	1.01	1.00	0.87
		64	2.43e+02	1.47	1.12	0.99	0.90	0.81	0.76	0.76	0.67
		128	3.02e+02	1.52	1.19	1.06	0.94	0.89	0.88	0.87	0.76
		256	5.56e+02	1.64	1.28	1.14	1.05	1.03	1.01	0.97	0.87
		512	1.06e+03	1.69	1.31	1.17	1.08	1.05	1.03	1.00	0.88
		1024	2.05e+03	1.72	1.31	1.18	1.09	1.04	1.04	1.00	0.88
		2048	4.06e+03	1.76	1.32	1.17	1.09	1.04	1.03	1.00	0.88
		4096	8.02e+03	1.72	1.30	1.15	1.07	1.03	1.01	0.99	0.87
		8192	1.60e+04	1.72	1.30	1.16	1.08	1.03	1.01	0.99	0.87
Qwen2.5-14B	Qwen2.5-14B	16384	3.17e+04	1.71	1.29	1.15	1.07	1.03	1.01	0.98	0.87
		64	2.74e+02	1.53	1.17	1.04	0.88	0.82	0.79	0.79	0.69
		128	3.47e+02	1.54	1.21	1.08	0.97	0.90	0.86	0.87	0.77
		256	6.39e+02	1.69	1.30	1.16	1.10	1.05	1.02	1.02	0.89
		512	1.23e+03	1.73	1.32	1.16	1.10	1.05	1.03	1.01	0.88
		1024	2.43e+03	1.76	1.33	1.19	1.13	1.07	1.05	1.03	0.89
		2048	4.80e+03	1.75	1.32	1.17	1.10	1.06	1.03	1.01	0.88
		4096	9.50e+03	1.75	1.31	1.16	1.09	1.04	1.01	1.00	0.87
		8192	1.88e+04	1.74	1.31	1.16	1.09	1.04	1.01	1.00	0.87
		16384	3.74e+04	1.72	1.30	1.15	1.08	1.03	1.01	0.99	0.87
GB10	Llama3.2-1B	64	2.06e+02	2.79	1.47	1.33	1.08	1.04	0.99	1.06	0.87
		128	2.28e+02	2.18	1.37	1.28	1.08	1.05	0.98	1.05	0.91
		256	2.96e+02	1.72	1.30	1.21	1.06	1.04	1.00	1.04	0.91
		512	4.10e+02	1.37	1.08	0.96	0.88	0.85	0.86	0.83	0.72
		1024	7.06e+02	1.22	0.96	0.85	0.80	0.75	0.75	0.74	0.64
		2048	1.43e+03	1.27	0.96	0.84	0.80	0.75	0.74	0.73	0.63
		4096	2.74e+03	1.23	0.93	0.82	0.77	0.72	0.72	0.70	0.62
		8192	5.56e+03	1.24	0.93	0.83	0.78	0.73	0.73	0.72	0.63
		16384	1.11e+04	1.22	0.93	0.83	0.77	0.74	0.73	0.71	0.63
		64	2.19e+02	2.38	1.35	1.23	1.14	0.99	0.95	1.02	0.84
BitNet-2B	BitNet-2B	128	2.49e+02	2.05	1.27	1.14	1.09	0.99	0.95	1.00	0.85
		256	3.20e+02	1.69	1.19	1.14	1.07	0.98	0.96	0.98	0.88
		512	4.52e+02	1.33	1.01	0.93	0.87	0.83	0.80	0.81	0.71
		1024	7.92e+02	1.24	0.92	0.85	0.79	0.77	0.74	0.72	0.64
		2048	1.52e+03	1.22	0.91	0.83	0.78	0.74	0.70	0.70	0.62

Appendix B Model Kernel Fp8 (Page 4/4)

GPU	Model	M	cuBLASLt Latency (μs)	cuSPARSELt Speedup Ratio							
				2:4	2:6	2:8	2:10	2:12	2:14	2:16	2:∞
		4096	3.07e+03	1.21	0.90	0.81	0.77	0.74	0.70	0.70	0.61
		8192	6.15e+03	1.21	0.90	0.81	0.76	0.74	0.70	0.70	0.61
		16384	1.24e+04	1.22	0.90	0.82	0.77	0.74	0.70	0.70	0.60
Llama3.2-3B	64	3.71e+02	2.35	1.62	1.36	1.12	1.14	0.95	1.10	0.93	
	128	4.00e+02	2.12	1.51	1.33	1.11	1.13	0.97	1.08	0.90	
	256	4.70e+02	1.80	1.40	1.23	1.03	1.06	0.95	1.04	0.90	
	512	6.64e+02	1.41	1.11	0.99	0.90	0.88	0.88	0.83	0.77	
	1024	1.09e+03	1.23	0.94	0.82	0.76	0.75	0.73	0.71	0.61	
	2048	2.31e+03	1.29	0.97	0.85	0.79	0.76	0.74	0.73	0.64	
	4096	4.45e+03	1.23	0.92	0.81	0.75	0.74	0.71	0.70	0.61	
	8192	9.11e+03	1.24	0.94	0.83	0.77	0.76	0.73	0.71	0.63	
Qwen2.5-7B	16384	1.83e+04	1.24	0.94	0.83	0.78	0.76	0.73	0.71	0.63	
	64	1.08e+03	1.97	1.26	1.27	1.14	0.92	1.02	1.05	0.91	
	128	1.15e+03	1.94	1.25	1.25	1.16	0.94	1.03	1.05	0.94	
	256	1.29e+03	1.85	1.27	1.23	1.15	0.96	1.05	1.05	0.94	
	512	1.61e+03	1.55	1.12	1.01	0.92	0.85	0.86	0.84	0.74	
	1024	2.78e+03	1.30	0.95	0.85	0.79	0.73	0.73	0.73	0.65	
	2048	5.58e+03	1.29	0.94	0.85	0.78	0.72	0.70	0.72	0.64	
	4096	1.12e+04	1.23	0.90	0.81	0.76	0.68	0.67	0.70	0.61	
Qwen2.5-14B	8192	2.25e+04	1.30	0.95	0.86	0.80	0.72	0.72	0.74	0.65	
	16384	4.45e+04	1.26	0.91	0.83	0.78	0.70	0.69	0.72	0.63	
	64	1.39e+03	2.36	1.42	1.36	1.25	1.04	0.96	1.13	0.94	
	128	1.44e+03	2.20	1.42	1.36	1.22	1.01	0.93	1.12	0.97	
	256	1.55e+03	1.99	1.38	1.31	1.19	1.03	0.95	1.11	0.92	
	512	1.97e+03	1.59	1.21	1.07	1.00	0.97	0.90	0.93	0.81	
	1024	3.12e+03	1.28	0.95	0.83	0.78	0.75	0.70	0.71	0.63	
	2048	6.21e+03	1.24	0.94	0.83	0.78	0.75	0.69	0.72	0.63	
Qwen2.5-14B	4096	1.26e+04	1.26	0.96	0.84	0.79	0.76	0.71	0.72	0.64	
	8192	2.48e+04	1.25	0.95	0.83	0.78	0.75	0.70	0.72	0.63	
	16384	5.00e+04	1.26	0.96	0.84	0.79	0.76	0.71	0.72	0.63	