	Model	Phi-3-mini-4k-instruct-q4.gguf											
	Prompt	What is an embedding model?"											
		load time (ms)	**sample time (ms)**	**runs**	**tokens per second**	**prompt eval time (ms)**	**tokens**	**tokens per second**	**eval time (ms)**	**runs**	**tokens per second**	**total time (ms)**	**tokens**
CPU	**12 cores**	119.17	27.23	256	9402.78	119.08	12	100.78	**12511.34**	255	20.38	**12975.24**	267
GPU	**RTX 4090**	130.02	25.53	256	10029.38	129.57	12	92.61	**1113.32**	255	229.05	**1524.88**	267
% delta		-9.10%	6.24%	0.00%	-6.66%	-8.81%	0.00%	8.11%	**91.10%**	0.00%	-1023.90%	**88.25%**	0.00%
							CPU	**Eval time / Total time**	96.42%			**GPU is about 8x faster**	
							GPU	**Eval time / Total time**	73.01%				
									Bulk of time spent in evaulation				
	Model	Phi-3-mini-4k-instruct-fp16.gguf											
	Prompt	What is an embedding model?"											
		load time (ms)	**sample time (ms)**	**runs**	**tokens per second**	**prompt eval time (ms)**	**tokens**	**tokens per second**	**eval time (ms)**	**runs**	**tokens per second**	**total time (ms)**	**tokens**
CPU	**12 cores**	229.3	17.72	168	9481.88	228.98	12	52.41	**23630.1**	167	7.07	**24105.49**	179
GPU	**RTX 4090**	136.41	24.9	256	10282.78	136.31	12	88.03	**2456**	255	103.83	**2866.99**	267
% delta		40.51%	-40.52%	-52.38%	-8.45%	40.47%	0.00%	-67.96%	**89.61%**	-52.69%	-1368.60%	**88.11%**	-49.16%
							CPU	**Eval time / Total time**	98.03%			**GPU is about 8x faster**	
							GPU	**Eval time / Total time**	85.66%				
									Bulk of time spent in evaulation				