Find a range of one-click templates	github.com/TrelisResearch/one-click-llms			
1	**H100 SXM**	**Llama 8B**		
Batch Size	**1**	**64**		
vLLM	130	25		
SGLang	156	130		
NIM	133	120		
TGI (fp8)	110	68		
TGI (bf16)	108	67		
llama.cpp	94	15	with -np 64	
1	**A40**	**Llama 8B**		
Batch Size	**1**	**64**		
llama.cpp	78	4		

Find a range of one-click templates at:		github.com/TrelisResearch/one-click-llms			**Setup**			
					SGLANG Inference			
					64	Batch Size		
Llama 8B	**toks**	**$/hr**	**$/mm output toks**		**Notes**			**toks batch=1**
A40 (INT4)	51	0.35	0.030		Not really worth it.			85
A40 (fp8)	43	0.35	0.035					
A6000 (fp8)	49	0.76	0.067					
A100 SXM (fp8)	115	1.94	0.073					
H100 SXM (fp8)	130	3.99	0.133					
Llama 70B	**toks**	**$/hr**	**$/mm output toks**		**Notes**			
2 x A40 (INT4)	16	0.35	0.190		INT4 is cheap as it fits you onto A40s...			24
2 x A40 (fp8)	OOM				toks are slow on cheaper hardware			
4 x A40 (INT4)	17	0.35	0.179					37
4 x A40 (fp8)	13.3	0.35	0.457					24
1 x A100 SXM (fp8)	OOM							
1 x H100 SXM (fp8)	7	3.99	2.474					30
2 x H100 SXM (fp8)	39	3.99	0.888					
4 x H100 SXM (fp8)	56	3.99	1.237					
GPT4o Mini ($0.15/mm input; $0.6/mm output)			0.6					
Llama 405B	**toks**	**$/hr**	**$/mm output toks**		**Notes**			
8 x A40 (INT4)	5.62	$0.34	2.101		toks are a bit too slow on cheap hardware			15
4 x H100 SXM (INT4)	17	$3.99	4.075				37	
8 x A100 SXM (fp8)	13.3	1.94	5.065					20
8 x H100 SXM (fp8)	25	3.99	5.542					32
GPT4o ($5/mm input; $15/mm output)			15.2					