Table 1: **Inference latency (ms) comparison across models, frameworks, and sequence lengths.** Tests are conducted on HuggingFace single-GPU deployment with MetaX C550 GPUs and on vLLM with both single- and multi-GPU settings. Latency is measured as the time to process the input sequence and generate 128 output tokens.

	SpikingBrain -7B	Mistral -7B-v0.1	Qwen2.5 -7B	SpikingBrain-7B	Mistral -7B-v0.1	Qwen2.5 -7B	**SpikingBrain -76B**	Mixtral -8x7B-v0.1	Llama -2-70b
GPUs	1			1			4		
Frameworks	HuggingFace			vllm			vllm		
Expert Parallel	Non-MoE						Enable / Disable		Non-MoE
32k	8452	10202	7969	4369	4570	6421	9451 / 5305	10477 / 3209	15881
64k	11685	16467	16257	7399	7084	15093	14077 / 8142	19062 / 4840	32953
128k	15974	28382	38487	12048	11867	45141	27106 / 14109	34983 / 8046	86120