Model	#Model Parameters (B)	Draft (Assistant)	#Draft Parameters (B)	Task	Total Parameter Size (B)	Speculative Average time per input (ms)	Speculative Average time per token (ms)	Original Average time per input (ms)	Original Average time per token (ms)	Speedup	Command
meta-llama/Llama-2-7b-hf	7	TinyLlama/TinyLlama_v1.1	1	summarization	8	2771.54	21.65	3368.48	26.32	1.22	python benchmark_decoder_summ.py meta-llama/Llama-2-7b-hf --aux-model TinyLlama/TinyLlama_v1.1 --dtype fp16
meta-llama/Llama-2-7b-hf	7	apple/OpenELM-270M	0.27	summarization	7.27	2607.82	20.37	4221.14	32.98	1.62	python benchmark_decoder_summ.py meta-llama/Llama-2-7b-hf --aux-model apple/OpenELM-270M --dtype fp16
meta-llama/Llama-2-7b-hf	7	apple/OpenELM-450M	0.45	summarization	7.45	3324.68	25.97	4178.66	32.65	1.26	python benchmark_decoder_summ.py meta-llama/Llama-2-7b-hf --aux-model apple/OpenELM-450M --dtype fp16
facebook/layerskip-llama2-7B	7	Early Exit @ Layer 4		summarization	7	2548.4	19.91	3306.73	25.83	1.297338021	python benchmark_decoder_summ.py facebook/layerskip-llama2-7B --aux-early-exit 4 --dtype fp16
meta-llama/Llama-2-13b-hf	13	meta-llama/Llama-2-7b-hf	7	summarization	20	3557.07	27.79	4088.48	31.94	1.149334293	python benchmark_decoder_summ.py meta-llama/Llama-2-13b-hf --aux-model meta-llama/Llama-2-7b-hf --dtype fp16
meta-llama/Llama-2-13b-hf	13	TinyLlama/TinyLlama_v1.1	1	summarization	14	2901.92	22.67	4190.42	32.74	1.444199382	python benchmark_decoder_summ.py meta-llama/Llama-2-13b-hf --aux-model TinyLlama/TinyLlama_v1.1 --dtype fp16
meta-llama/Llama-2-13b-hf	13	apple/OpenELM-270M	0.27	summarization	13.27	2883.33	22.53	4521.12	35.32	1.567687528	python benchmark_decoder_summ.py meta-llama/Llama-2-13b-hf --aux-model apple/OpenELM-270M --dtype fp16
meta-llama/Llama-2-13b-hf	13	apple/OpenELM-450M	0.45	summarization	13.45	3267.69	25.53	4321.75	33.76	1.322365844	python benchmark_decoder_summ.py meta-llama/Llama-2-13b-hf --aux-model apple/OpenELM-450M --dtype fp16
facebook/layerskip-llama2-13B	13	Early Exit @ Layer 4		summarization	13	4238.45	33.11	4217.78	32.95	0.9951676231	python benchmark_decoder_summ.py facebook/layerskip-llama2-13B --aux-early-exit 4 --dtype fp16
facebook/layerskip-llama2-13B	13	Early Exit @ Layer 8		summarization	13	2459.61	19.22	4294.98	33.55	1.745577523	python benchmark_decoder_summ.py facebook/layerskip-llama2-13B --aux-early-exit 8 --dtype fp16
facebook/layerskip-llama3.2-1B	1	Early Exit @ Layer 4		summarization	1	1195.28	9.96	2147.7	17.9	1.80	python benchmark_decoder_summ.py facebook/layerskip-llama3.2-1B --aux-early-exit 4 --dtype fp16
meta-llama/Meta-Llama-3-8B	8	meta-llama/Llama-3.2-1B	1	summarization	9	1872.46	19.04	2859.35	29.08	1.53	python benchmark_decoder_summ.py meta-llama/Meta-Llama-3-8B --aux-model meta-llama/Llama-3.2-1B --dtype fp16
meta-llama/Meta-Llama-3-8B	8	meta-llama/Llama-3.2-3B	3	summarization	11	2814.82	28.63	2825.36	28.73	1.00	python benchmark_decoder_summ.py meta-llama/Meta-Llama-3-8B --aux-model meta-llama/Llama-3.2-3B --dtype fp16
facebook/layerskip-llama3-8B	8	Early Exit @ Layer 4		summarization	8	1949.02	15.75	3571.81	28.87	1.83	python benchmark_decoder_summ.py facebook/layerskip-llama3-8B --aux-early-exit 4 --dtype fp16
meta-llama/Llama-2-70b-hf	70	meta-llama/Llama-2-13b-hf	13	summarization	83	5036.54	46.3	12289.01	112.97	2.439956803	python benchmark_decoder_summ.py meta-llama/Llama-2-70b-hf --aux-model meta-llama/Llama-2-13b-hf --dtype fp16
meta-llama/Llama-2-70b-hf	70	meta-llama/Llama-2-7b-hf	7	summarization	77	4357.55	40.06	12324.19	113.3	2.828257614	python benchmark_decoder_summ.py meta-llama/Llama-2-70b-hf --aux-model meta-llama/Llama-2-7b-hf --dtype fp16
meta-llama/Llama-2-70b-hf	70	TinyLlama/TinyLlama_v1.1	1	summarization	71	4356.21	40.05	12363.22	113.66	2.837952559	python benchmark_decoder_summ.py meta-llama/Llama-2-70b-hf --aux-model TinyLlama/TinyLlama_v1.1 --dtype fp16
facebook/layerskip-llama2-70B	70	Early Exit @ Layer 10		summarization	70	6012.04	54.96	12383.34	113.2	2.06	python benchmark_decoder_summ.py facebook/layerskip-llama2-70B --aux-early-exit 10 --dtype fp16