	# Params	# Layers	# Hidden size	max length	model_type	pipe_mode	reference	Layer-wise Checkpoint	Comment
T5	11B	24	1024	512 (prompt) + 512 (decode)	t5	pipe_async_sample_enc_dec_mask	https://huggingface.co/t5-11b	https://pretrained-models-inference.s3.(	encoder-decoder model
T0pp	11B	24	1024	512 (prompt) + 512 (decode)	t5	pipe_async_sample_enc_dec_mask	https://huggingface.co/bigscience/T0pp	https://pretrained-models-inference.s3.(	encoder-decoder model
UL2	20B	32	4096	512 (prompt) + 512 (decode)	t5	pipe_async_sample_enc_dec_mask	https://huggingface.co/google/ul2	https://pretrained-models-inference.s3.(	encoder-decoder model
GPT-2	110M	12	768	1024	gpt2	pipe_sync_sample_mask_token_pipe	https://huggingface.co/gpt2	-	decoder only
GPT-J	6B	28	4096	2048	gptj	pipe_sync_sample_mask_token_pipe	https://huggingface.co/EleutherAI/gpt-j-6B	-	decoder only
GPT-NeoX	20B	44	6144	2048	gptneox	pipe_sync_sample_mask_token_pipe	https://huggingface.co/EleutherAI/gpt-neox-20	https://pretrained-models-inference.s3.(	decoder only
YaLM	100B	80	10240	2048	yalm	<not in the current version>	https://github.com/yandex/YaLM-100B/	https://pretrained-models-inference.s3.(	decoder only (this is a reproduction)
OPT-66B	66B	64	9216	2048	opt	pipe_sync_sample_mask_token_pipe	https://huggingface.co/facebook/opt-66b	https://pretrained-models-inference.s3.(	decoder only
OPT-175B	175B	96	12288	2048	opt	pipe_sync_sample_mask_token_pipe	not publicly available	-	decoder only
Bloom	175B	70	14336	2048	bloom	pipe_sync_sample_mask_token_pipe	https://huggingface.co/bigscience/bloom	https://pretrained-models-inference.s3.(	decoder only