

	# 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	<a href="https://huggingface.co/t5-11b">https://huggingface.co/t5-11b</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/t5-11b/layer_wise_checkpoint/t5-11b-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/t5-11b/layer_wise_checkpoint/t5-11b-layer_wise_checkpoint.tar.gz</a>	encoder-decoder model
T0pp	11B	24	1024	512 (prompt) + 512 (decode)	t5	pipe_async_sample_enc_dec_mask	<a href="https://huggingface.co/bigscience/T0pp">https://huggingface.co/bigscience/T0pp</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/t0pp/layer_wise_checkpoint/t0pp-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/t0pp/layer_wise_checkpoint/t0pp-layer_wise_checkpoint.tar.gz</a>	encoder-decoder model
UL2	20B	32	4096	512 (prompt) + 512 (decode)	t5	pipe_async_sample_enc_dec_mask	<a href="https://huggingface.co/google/ul2">https://huggingface.co/google/ul2</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/ul2/layer_wise_checkpoint/ul2-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/ul2/layer_wise_checkpoint/ul2-layer_wise_checkpoint.tar.gz</a>	encoder-decoder model
GPT-2	110M	12	768	1024	gpt2	pipe_sync_sample_mask_token_pipe	<a href="https://huggingface.co/gpt2">https://huggingface.co/gpt2</a>	-	decoder only
GPT-J	6B	28	4096	2048	gptj	pipe_sync_sample_mask_token_pipe	<a href="https://huggingface.co/EleutherAI/gpt-j-6B">https://huggingface.co/EleutherAI/gpt-j-6B</a>	-	decoder only
GPT-NeoX	20B	44	6144	2048	gptneox	pipe_sync_sample_mask_token_pipe	<a href="https://huggingface.co/EleutherAI/gpt-neox-20b">https://huggingface.co/EleutherAI/gpt-neox-20b</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/gpt-neox-20b/layer_wise_checkpoint/gpt-neox-20b-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/gpt-neox-20b/layer_wise_checkpoint/gpt-neox-20b-layer_wise_checkpoint.tar.gz</a>	decoder only
YaLM	100B	80	10240	2048	yalm	<not in the current version>	<a href="https://github.com/yandex/YaLM-100B/">https://github.com/yandex/YaLM-100B/</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/yalm/layer_wise_checkpoint/yalm-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/yalm/layer_wise_checkpoint/yalm-layer_wise_checkpoint.tar.gz</a>	decoder only (this is a reproduction)
OPT-66B	66B	64	9216	2048	opt	pipe_sync_sample_mask_token_pipe	<a href="https://huggingface.co/facebook/opt-66b">https://huggingface.co/facebook/opt-66b</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/opt-66b/layer_wise_checkpoint/opt-66b-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/opt-66b/layer_wise_checkpoint/opt-66b-layer_wise_checkpoint.tar.gz</a>	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	<a href="https://huggingface.co/bigscience/bloom">https://huggingface.co/bigscience/bloom</a>	<a href="https://pretrained-models-inference.s3.amazonaws.com/bloom/layer_wise_checkpoint/bloom-layer_wise_checkpoint.tar.gz">https://pretrained-models-inference.s3.amazonaws.com/bloom/layer_wise_checkpoint/bloom-layer_wise_checkpoint.tar.gz</a>	decoder only