A TECHNICAL APPENDIX

A.1 Statistic of Datasets

Statistics of the GLUE benchmark and the generation datasets are shown in Table 1. The first eight datasets are from the GLUE benchmark, and the rest are generation datasets.

Table 1: Statistics of the datasets. WMT is short for WMT 2016 en-ro and OBQA is short for OpenBookQA.

Tasks	MNLI	QQP	QNLI	SST-2	CoLA
#Train	392.7k	363.8k	104.7k	67.4k	8.6k
#Dev.	9.8k	40.4k	5.5k	0.9k	1k
Tasks	RTE	STS-B	MRPC		
#Train	2.5k	5.8k	3.7k		
#Dev.	0.3k	1.5k	0.4k		
Tasks	WMT	OBQA	XSum	Alpaca	MMLU
#Train	610.3k	5k	204.0k	52k	-
#Dev.	2.0k	0.5k	11.3k	-	-
#Test	2.0k	0.5k	11.3k	-	14k

A.2 Details of the Implementation

A.2.1 Hyperparameters on the GLUE Benchmark. Table 2 shows the hyperparameters used in the GLUE benchmark experiments. For the first stage which aims to learn the parameter masks, we set the ranks of the low-rank weight matrices to 4 and 3 for RoBERTa-large and RoBERTa-base respectively for all datasets so that the ratio of trainable parameters is about 0.5% for fair comparison. We run all the experiments on GLUE using a single NVIDIA RTX4090 GPU with 24GB memory. In addition, for all experiments in this paper, we use AdamW optimizer with a warm up ratio of 0.06 and a linear learning rate schedule.

Table 2: The hyper-parameters of LoReML on the GLUE benchmark using RoBERTa-large and RoBERTa-base as backbone models. Seq. Len. denotes sequence length.

HP	MNLI	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE
First stage (RoBERTa-large/RoBERTa-base)								
Batch size	32	64/32	32/16	32	32/64	32	32	32
Epochs	10/20	20	10/15	10/5	40/80	20/40	20/30	20
Learning rate	3e-5/2e-4	1e-4/2e-4	5e-5/3e-5	5e-5	2e-4/4e-4	5e-5	5e-5/1e-4	5e-5/1e-4
rank r	4/3							
α	8/6	8/6	8/3	8/6	4/3	8/6	8/3	8/6
Seq. Len.	128/512	128/512	128	128	128	128	128/512	512
Second stage (RoBERTa-large/RoBERTa-base)								
Batch size	32/128	32/64	16/64	32/16	32/16	8/16	8	8
Epochs	10/35	15/40	20/35	20/10	30/80	15/25	15/35	20/40
Learning rate	1e-4	3e-4/2e-4	2e-4/1e-4	1e-4/2e-5	2e-4/4e-4	4e-4/5e-4	2e-4/1e-4	3e-4/2e-4
α	4/1.5	0/1.5	2/1.5	2/6	4/0	0/6	8/1.5	8/1.5
Seq. Len.	128	128	128	128	128	128	512	512

A.2.2 Hyperparameters on the Generation Tasks. The setting of the ranks for mask learning is according to the ratio of trainable parameters for fair comparison. Specifically, the ranks in mask learning are 40, 58, 16 and 16 for BART-large, mBART-large, OPT-7B and LLaMA-7B, respectively, and α in mask learning is equal to the rank for different models. For the second stage, α are set to 20 and 29 for BART-large and mBART-large, respectively, 8 for XSum dataset on OPT-7B and LLaMA-7B, 0 and 16 for OpenBookQA on OPT-7B and LLaMA-7B respectively. The optimizer setting is the same as that in the GLUE experiments. The remaining hyper-parameters for the generation tasks are shown in Table 3.

Table 3: Hyperparameters for the generation tasks.

Model	HP	XSum	WMT 2016	OpenBookQA
First stage/Second stage				
BART-large	Batch size	32	32	-
	Epochs	10	5/2	-
	Learning rate	1e-4	1e-4/1e-5	-
	Seq. Len.	512	150	-
OPT-7B	Batch size	32	-	32
	Epochs	5	-	8
	Learning rate	5e-5	-	5e-5
	Seq. Len.	320	-	216
Model	HP	XSum	MMLU	OpenBookQA
LLaMA-7B	Batch size	16/32	16	16/32
	Epochs	5	3	8
	Learning rate	5e-5	1e-4	5e-5/1e-4
	Seq. Len.	320	512	216
LLaMA-13B	Batch size	16/32	16	16/32
	Epochs	3/5	3	8
	Learning rate	1e-4/2e-4	1e-4	5e-5/3e-4
	Seq. Len.	320	512	216