Table 1: We fine-tune Llama2-7B on the StrategyQA dataset using supervised fine-tuning (SFT) and selectively supervised fine-tuning (SSFT). Here are the capabilities of models on four commonsense reasoning tasks (*e.g.*, Winogrande, CSQA, and StrategyQA) before and after tuning.

		ID Task		**OOD Task**					
		StrategyQA		**CSQA**		**Winogrande**		**Average**	
Models	**Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Llama2-7B	-	62.5	-	61.1	-	53.4	-	57.3	-
+ SFT	6.7B	77.3	+14.8	54.8	-6.3	52.7	-0.7	53.8	-3.5
+ SSFT	0.2B	78.5	+16.0	64.1	+3.0	61.1	+7.7	62.6	+5.3

Table 2: We fine-tune Llama2-7B on the CSQA dataset using supervised fine-tuning (SFT) and selectively supervised fine-tuning (SSFT). Here are the capabilities of models on four commonsense reasoning tasks (*e.g.*, Winogrande, CSQA, and StrategyQA) before and after tuning.

		ID Task		**OOD Task**					
		CSQA		**StrategyQA**		**Winogrande**		**Average**	
Models	**Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Llama2-7B	-	61.1	-	62.5	-	53.4	-	58.0	-
+ SFT	6.7B	72.3	+11.2	57.8	-4.7	53.5	+0.1	55.7	-0.3
+ SSFT	0.2B	73.5	+12.4	63.1	+0.6	56.2	+2.8	59.7	+1.7