Table 1: We fine-tune Llama2-7B/13B on the Winogrande 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, StrategyQA, and SocialIQA) before and after tuning.

		ID Task		**OOD Task**							
		Winogrande		**CSQA**		**StrategyQA**		**SocialIQA**		**Average**	
Models	**Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Llama2-7B	-	53.4	-	61.1	-	62.5	-	60.2	-	61.3	-
+ SFT	6.7B	74.3	+20.9	56.8	-4.3	64.3	+1.8	59.0	-1.2	60.0	-1.3
+ SSFT	0.2B	74.3	+20.9	64.9	+3.8	63.2	+0.7	63.2	+3.0	63.8	+2.5
Llama2-13B	-	55.5	-	68.3	-	66.0	-	67.9	-	63.9	-
+ SFT	13B	77.3	+21.8	64.6	-3.7	69.9	+3.9	63.1	-4.8	65.9	+2.0
+ SSFT	0.5B	75.6	+20.1	71.7	+3.4	68.6	+2.6	69.2	+1.3	69.8	+5.9

Table 2: We fine-tune Llama2-7B/13B on the SocialIQA 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.*, SocialIQA, CSQA, StrategyQA, and Winogrande) before and after tuning.

		ID Task		**OOD Task**							
		SocialIQA		**CSQA**		**StrategyQA**		**Winogrande**		**Average**	
Models	**Tuned Params.**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**	**Acc.**	**Δ**
Llama2-7B	-	60.2	-	61.1	-	62.5	-	53.4	-	59.0	-
+ SFT	6.7B	74.5	+14.3	65.9	+4.8	63.0	0.5	52.6	-0.8	60.5	+1.5
+ SSFT	0.2B	75.1	+14.9	66.4	+5.3	65.2	+2.7	55.8	+2.4	62.5	+3.5
Llama2-13B	-	67.9	-	68.3	-	66.0	-	55.5	-	63.3	-
+ SFT	13B	74.7	+6.8	67.2	-1.1	66.4	+0.4	53.5	-2.0	62.4	-0.9
+ SSFT	0.5B	75.1	+7.2	70.7	+2.4	69.9	+3.9	55.8	+0.3	65.5	+2.2