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

| | | **ID Task** | | **OOD Task** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **StrategyQA** | | **CSQA** | | **Winogrande** | | **SocialIQA** | | **Average** | |
| **Models** | **Tuned Params.** | **Acc.** | **Δ** | **Acc.** | **Δ** | **Acc.** | **Δ** | **Acc.** | **Δ** | **Acc.** | **Δ** |
| Llama2-7B | - | 62.5 | - | 61.1 | - | 53.4 | - | 60.2 | - | 58.2 | - |
| + SFT | 6.7B | 77.3 | +14.8 | 54.8 | -6.3 | 52.7 | -0.7 | 59.0 | -1.2 | 55.5 | -2.7 |
| + SSFT | 0.2B | 78.5 | +16.0 | 64.1 | +3.0 | 61.1 | +7.7 | 63.2 | +3.1 | 62.8 | +4.6 |
| Llama2-13B | - | 66.0 | - | 68.3 | - | 55.5 | - | 67.9 | - | 63.9 | - |
| + SFT | 13B | 79.0 | +13.0 | 69.5 | +1.2 | 54.6 | -0.9 | 63.1 | -4.8 | 62.4 | -1.5 |
| + SSFT | 0.5B | 80.3 | +14.3 | 72.6 | +4.3 | 56.6 | +1.1 | 69.2 | +1.3 | 66.1 | +2.2 |