Finetuning dataset	**MMLU**	**HellaSwag**	**GSM8K**	**TriviaQA**	**NQ-Open**
Original	53.42	56.31	34.65	47.63	11.61
sd (ep2)→msd (ep2)	53.28 (−0.14)	56.21 (−0.10)	33.78 (−0.87)	47.81 (+0.18)	11.82 (+0.21)
sd (ep2)→sdvar (ep2)	53.16 (−0.26)	56.15 (−0.16)	33.72 (−0.93)	47.60 (−0.03)	11.89 (+0.28)
IN2 (ep2)→IN2 (ep2)	53.45 (+0.03)	56.36 (+0.05)	34.25 (−0.40)	44.72 (−2.91)	9.58 (−2.03)
MultidocQA (ep2)→MultidocQA (ep2)	53.24 (−0.18)	56.22 (−0.09)	31.77 (−2.88)	44.80 (−2.83)	9.36 (−2.25)

Table 7: Mistral 7B and finetuned versions' performance evaluated on general ability benchmarks. All numbers are reported in percentage.