Finetuning dataset	MMLU	HellaSwag	GSM8K	TriviaQA	NQ-Open
Original	53.42	56.31	34.65	47.63	11.61
MultidocQA (ep1)	53.16 (−0.26)	56.16 (−0.15)	34.08 (−0.57)	45.70 (−1.93)	8.57 (−3.04)
MultidocQA (ep2)	53.19 (−0.22)	56.27 (−0.04)	33.28 (−1.36)	45.20 (−2.43)	8.69 (−2.91)
MultidocQA (ep4)	53.19 (−0.23)	56.37 (+0.06)	33.05 (−1.60)	44.93 (−2.70)	7.63 (−3.98)
MultidocQA x2 (ep2)	52.89 (−0.53)	56.20 (−0.11)	33.00 (−1.65)	44.77 (−2.86)	8.15 (−3.46)

Table 4: Mistral 7B and finetuned versions' performance evaluated on general ability benchmarks. All numbers are reported in percentage.