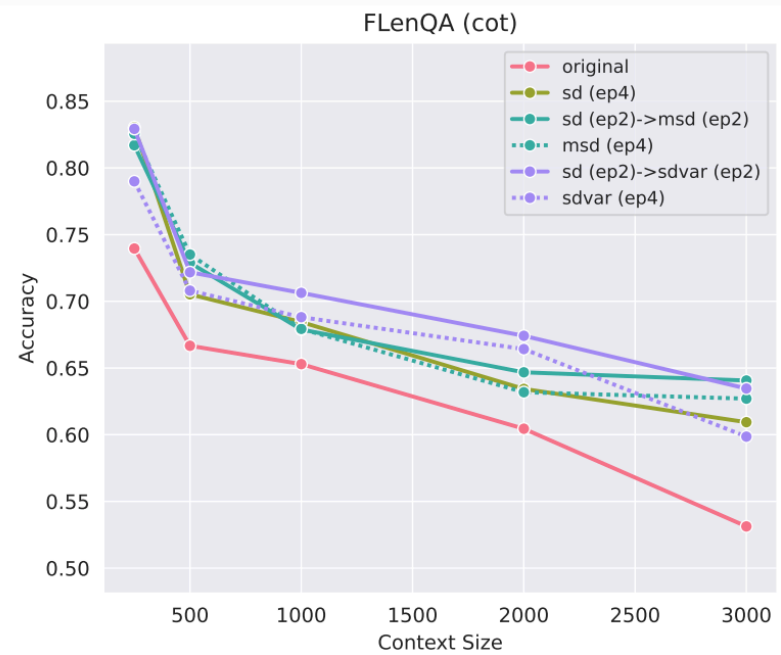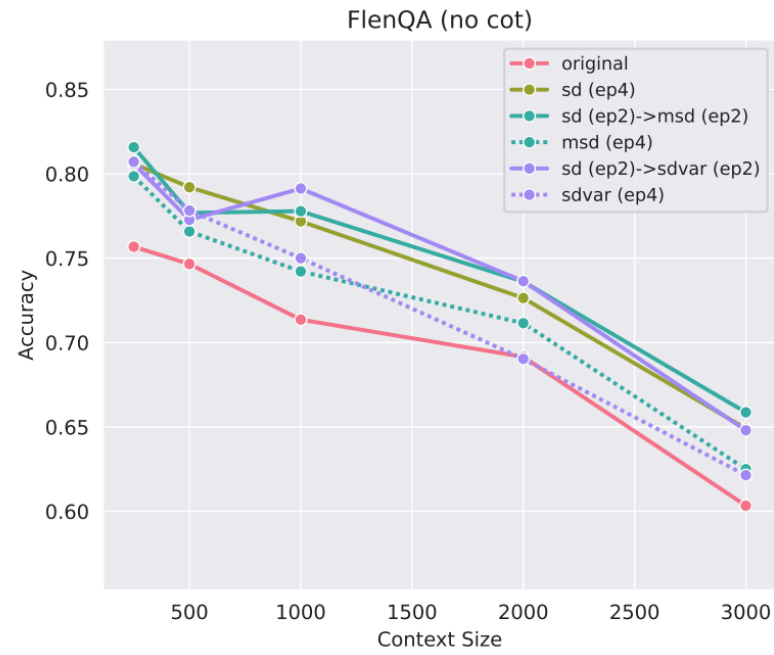(a) MDQA

(b) FLenQA with chain-of-thought prompting

(c) FLenQA without chain-of-though prompting

Figure I: Performance of finetuned Mistral 7B.