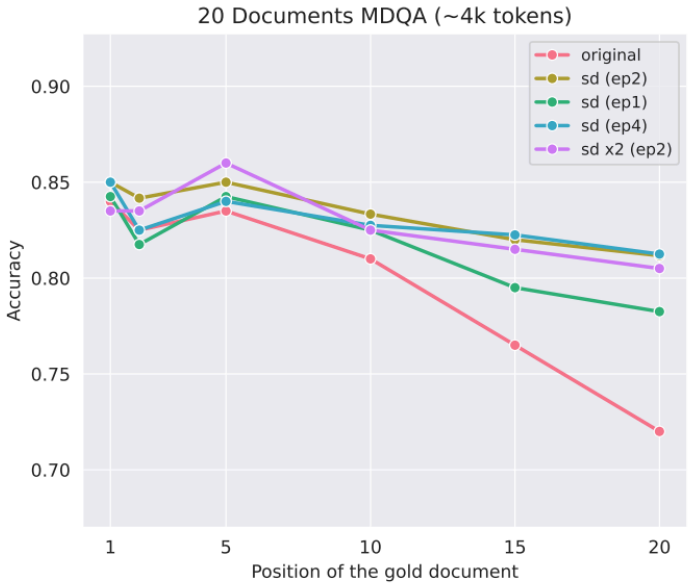
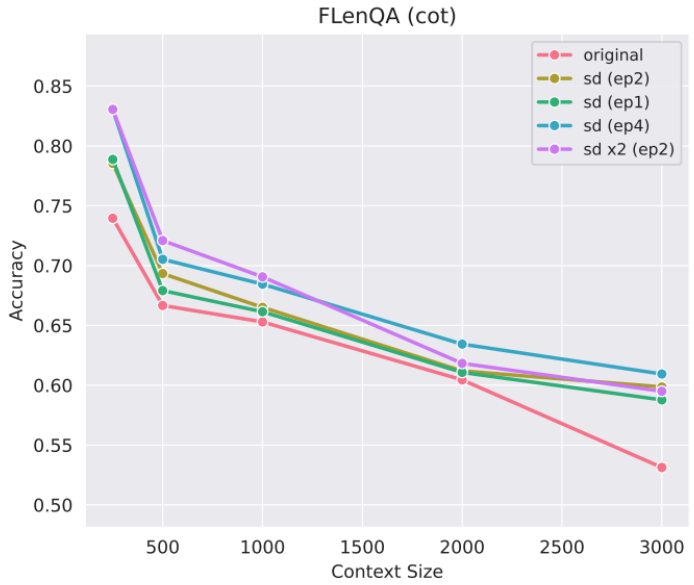
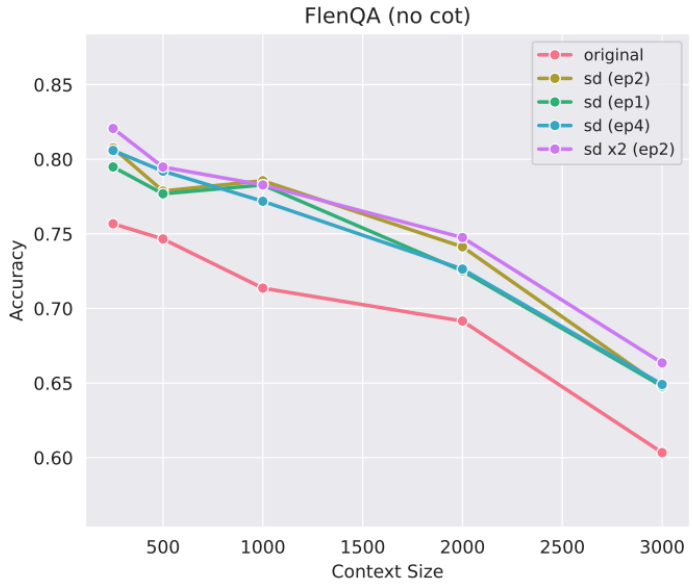


(a) MDQA



(b) FLenQA with chain-of-thought prompting



(c) FLenQA without chain-of-though prompting

Figure 12: Performance of finetuned Mistral 7B with different training epochs and training sizes, e.g., "sd (ep2)" denotes training on simple dictionary key-value retrieval task (sd) with 2 epochs; "sd x2 (ep2)" denotes training on sd task with 2 epochs but with training data twice as large. Subplots show the average performance of (a) MDQA, (b) FLenQA with chain-of-thought prompting, and (c) FLenQA without chain-of-thought prompting.