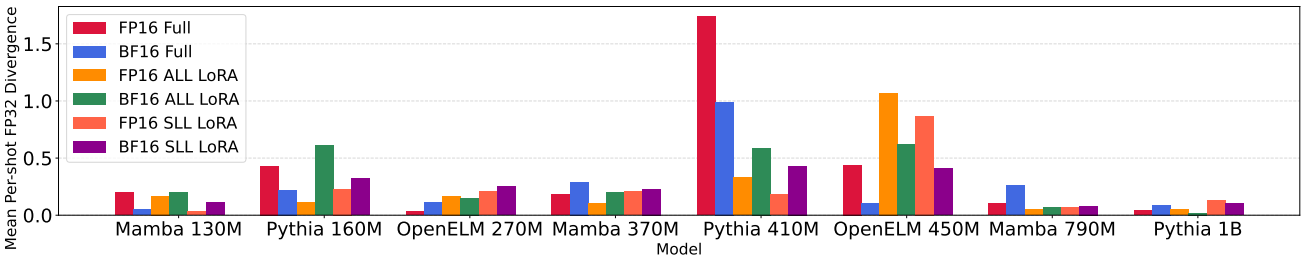
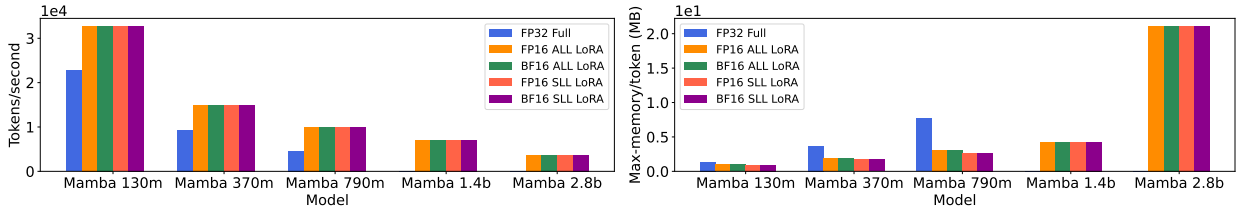


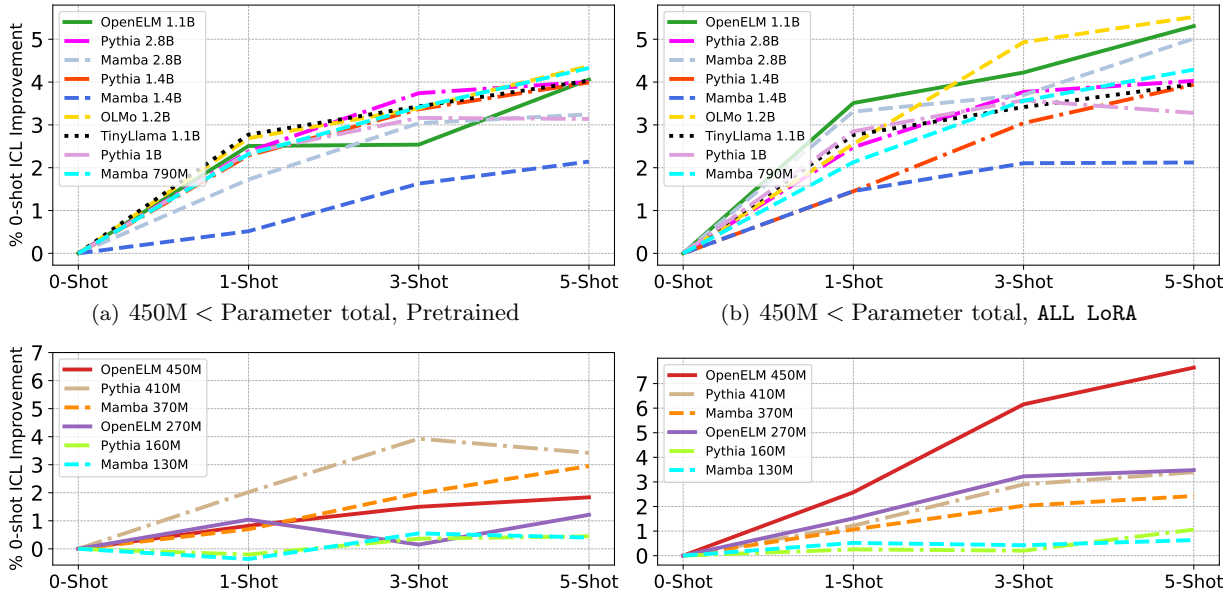
Figure 1: Mean full-precision (`FP32`) divergence in MMLU performance for Mamba, Pythia, and OpenELM models. Models are fine-tuned over the `Alpaca` dataset using different combinations of MPFT and PEFT. Full fine-tuning (i.e., no PEFT adapters) is denoted as `Full`.



(a) Average tokens-per-second ↑.

(b) Maximum memory-per-token ↓.

Figure 2: Timing and memory usage calculated Mamba model-sizes and PEFT combinations. Each model was trained using the Alpaca dataset dataset for three epochs and maximum sequence length 512. For each PEFT combination, the batch size was tuned to maximize GPU occupancy.



(a) 450M < Parameter total, Pretrained

(b) 450M < Parameter total, `ALL LoRA`

(c) Parameter total ≤ 450M, Pretrained

(d) Parameter total ≤ 450M, `ALL LoRA`

Figure 3: Fine-tuning improves Mamba ICL performance towards that of Transformer LLMs. `ALL LoRA` models were instruction fine-tuned on the OpenHermes dataset for one epoch. Performance is reported as the average improvement percentage of {1, 3, 5}-shot versus 0-shot over five standard natural language benchmarks: HellaSwag, PIQA, Arc-E, Arc-C, and WinoGrande.