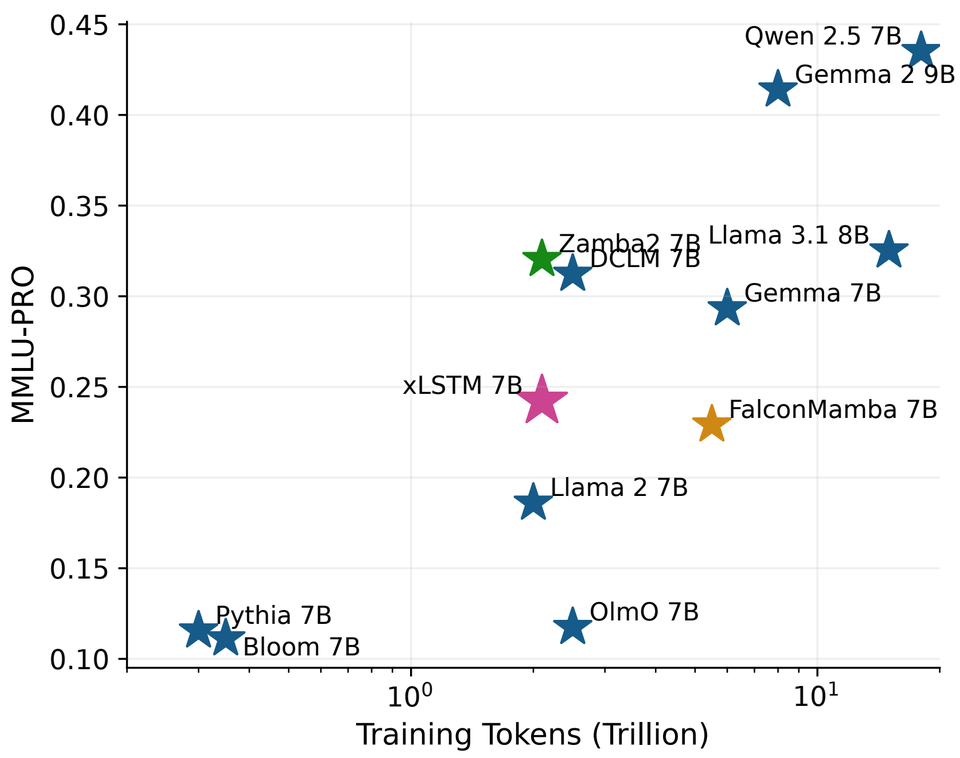
xLSTM 7B

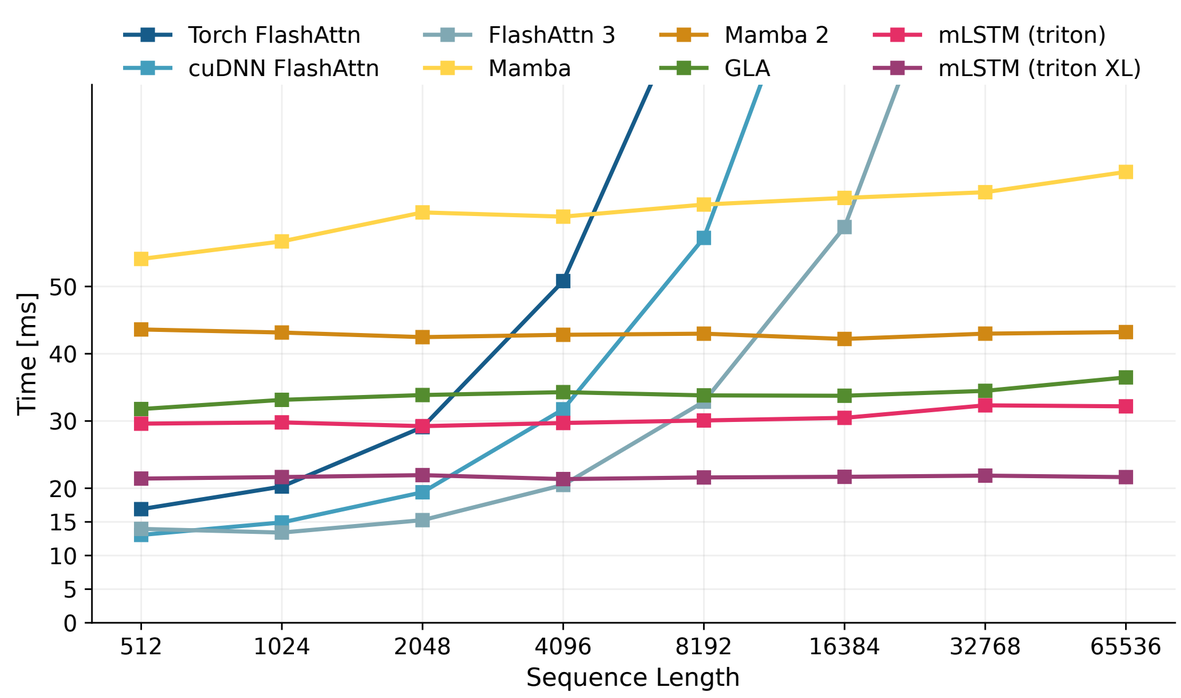
NXAI xLSTM Team: Maximilian Beck, Korbinian Pöppel, Phillip Lippe,
Richard Kurle, Patrick Blies, Sebastian Böck and Sepp Hochreiter

Downstream Performance



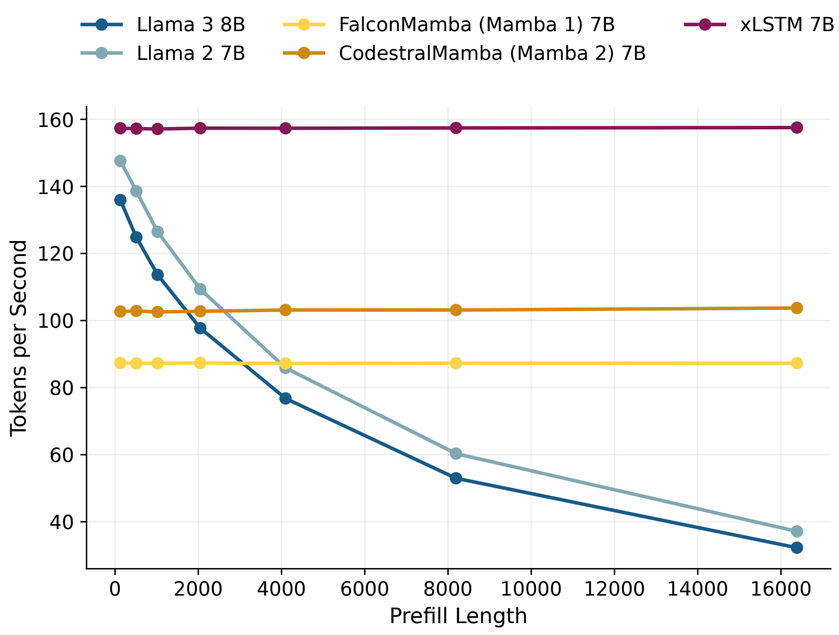
Downstream performance on MMLU-PRO

mLSTM Kernel Speed



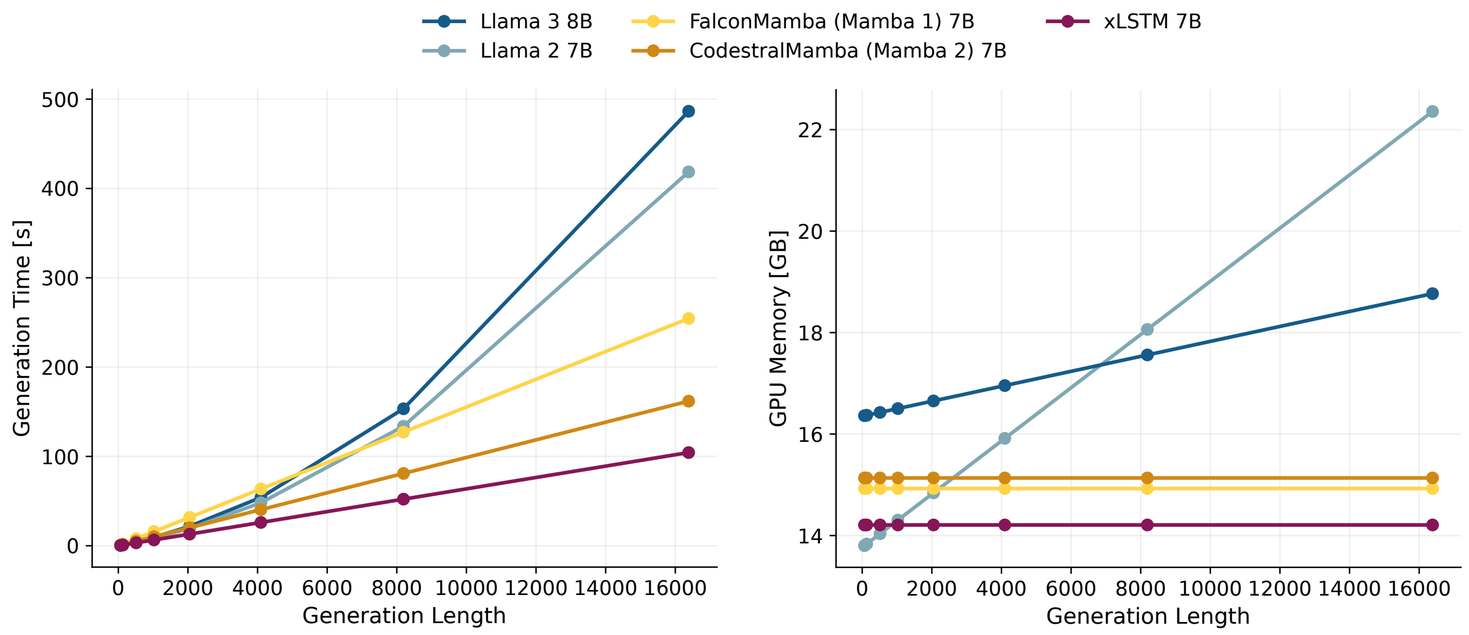
Forward+Backward kernel runtime for varying sequence length for 65k tokens with embedding dimension 4096

Generation Throughput



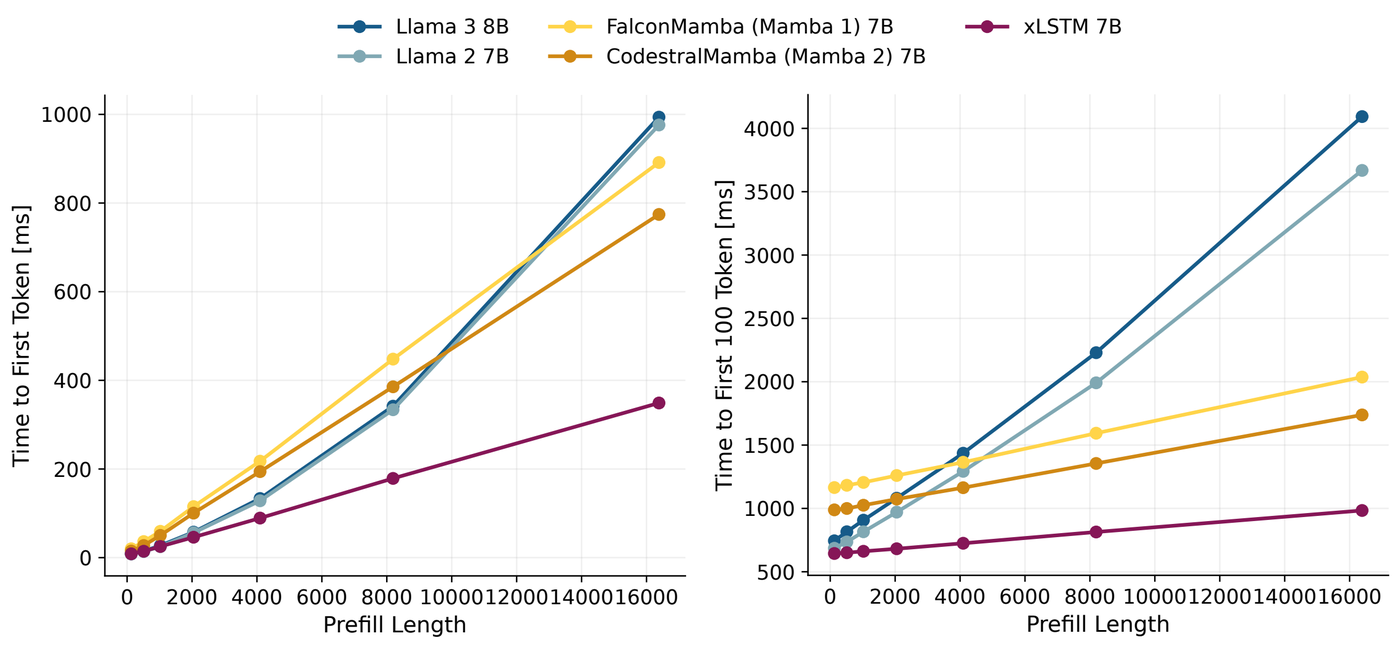
Throughput for generating of 100 tokens with batch size 1 at varying prefill lengths

Generation Times and Memory Consumption



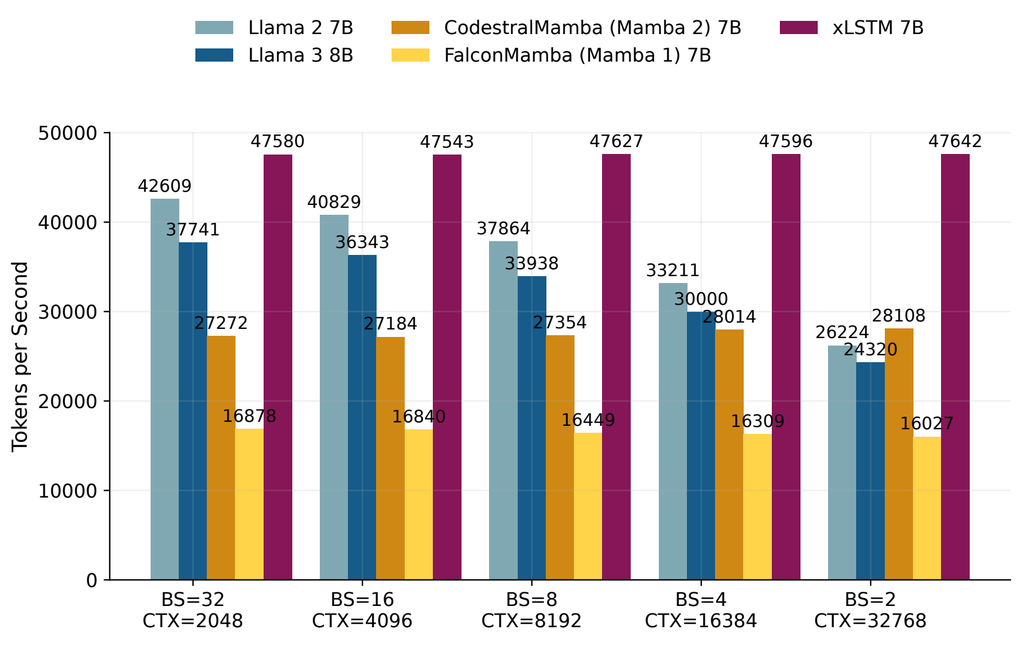
Time and GPU memory used for generation of a single sequence of varying lengths. Generation without prefill.

Time To First Token



Time to first and time to first 100 tokens at varying prefill lengths for batch size 1.

Throughput



Forward Throughput for varying batch sizes and context lengths

All speed experiments were run on NVIDIA H100 GPUs. We use implementations from Huggingface transformers library and optimize with torch.compile() and PyTorch CUDA Graphs.