xLSTM
Extended Long Short-Term Memory

Maximilian Beck*[1,2,3], Korbinian Pöppel*[1,2,3], Markus Spanring[1], Andreas Auer[1,2], Oleksandra Prudnikova
Michael Kopp, Günter Klambauer[1,2,3], Johannes Brandstetter[1,2,3], Sepp Hochreiter[1,2,3]
1) Johannes Kepler University Linz, Institute for Machine Learning 2) NXAI Lab Linz 3) NXAI GmbH

ellis European Laboratory for Learning and Intelligent Systems

Can LSTMs be scaled to billions of parameters while matching Transformer's capabilities?

Limitations of the LSTM:

LSTM's inability to revise storage decisions

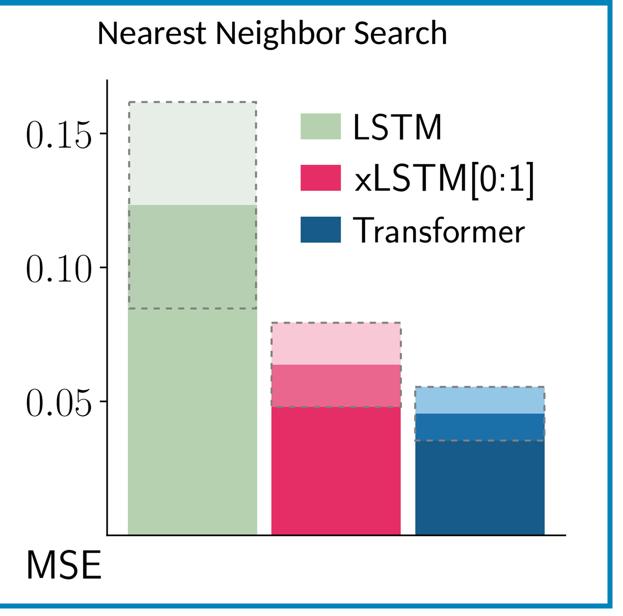
- Sigmoid input gate is limited → cannot overwrite
- Replace by *exponential input gate*
- Introduce normalizer $\boldsymbol{n}_t$ to re-stabilize

LSTM

$$\begin{aligned} \boldsymbol{c}_t &= \sigma(\tilde{\mathbf{f}}_t) \odot \boldsymbol{c}_{t-1} + \sigma(\tilde{\mathbf{i}}_t) \odot \tanh(\tilde{\boldsymbol{z}}_t) \\ \boldsymbol{h}_t &= \sigma(\tilde{\mathbf{o}}_t) \odot \tanh(\boldsymbol{c}_t) \end{aligned}$$

sLSTM

$$\begin{aligned} \boldsymbol{c}_t &= \sigma(\tilde{\mathbf{f}}_t) \odot \boldsymbol{c}_{t-1} + \exp(\tilde{\mathbf{i}}_t) \odot \tanh(\tilde{\boldsymbol{z}}_t) \\ \boldsymbol{n}_t &= \sigma(\tilde{\mathbf{f}}_t) \odot \boldsymbol{n}_{t-1} + \exp(\tilde{\mathbf{i}}_t) \\ \boldsymbol{h}_t &= \sigma(\tilde{\mathbf{o}}_t) \odot \boldsymbol{c}_t \,/\, \boldsymbol{n}_t \end{aligned}$$



LSTM's limited storage capacity

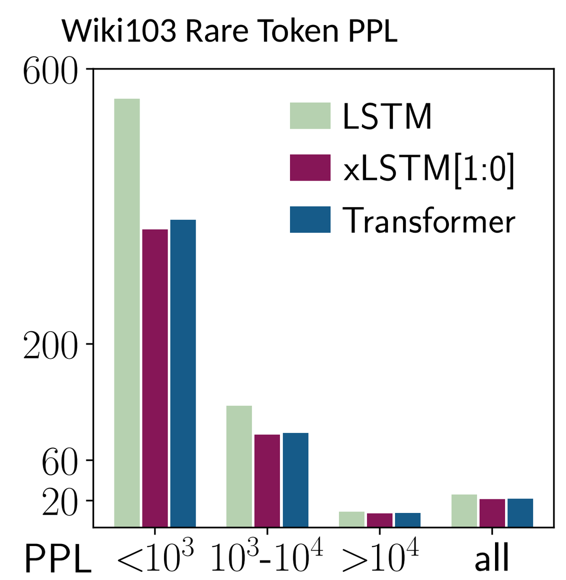
- Scalar memory cells, each gated → limited capacity
- Now *matrix memory cell* with *outer product update*
- Use down-projection to hidden state by query vector

LSTM

$$\begin{aligned} \boldsymbol{c}_t &= \sigma(\tilde{\mathbf{f}}_t) \odot \boldsymbol{c}_{t-1} + \sigma(\tilde{\mathbf{i}}_t) \odot \tanh(\tilde{\boldsymbol{z}}_t) \\ \boldsymbol{h}_t &= \sigma(\tilde{\mathbf{o}}_t) \odot \tanh(\boldsymbol{c}_t) \end{aligned}$$

mLSTM

$$\begin{aligned} \boldsymbol{C}_t &= \sigma(\tilde{\mathrm{f}}_t) \odot \boldsymbol{C}_{t-1} + \exp(\tilde{\mathrm{i}}_t) \odot \boldsymbol{v}_t \boldsymbol{k}_t^\top \\ \boldsymbol{n}_t &= \sigma(\tilde{\mathrm{f}}_t) \odot \boldsymbol{n}_{t-1} + \exp(\tilde{\mathrm{i}}_t) \odot \boldsymbol{k}_t \\ \boldsymbol{h}_t &= \mathbf{o}_t \odot \boldsymbol{C}_t \boldsymbol{q}_t \,/\, \max(|\boldsymbol{n}_t^\top \boldsymbol{q}_t|, 1) \end{aligned}$$



LSTM's inability of time-parallel training

- Recurrent connection limits parallelizability
- *Remove recurrent connection*
- Or: Block-diagonal / multi-head structure for Recurrent Matrix

LSTM

$$\{\tilde{\mathbf{i}}, \tilde{\mathbf{f}}, \tilde{\mathbf{z}}, \tilde{\mathbf{o}}\}_t = \boldsymbol{W}_{\{\mathrm{i,f,z,o}\}} \boldsymbol{x}_t + \boldsymbol{R}_{\{\mathrm{i,f,z,o}\}} \boldsymbol{h}_{t-1}$$

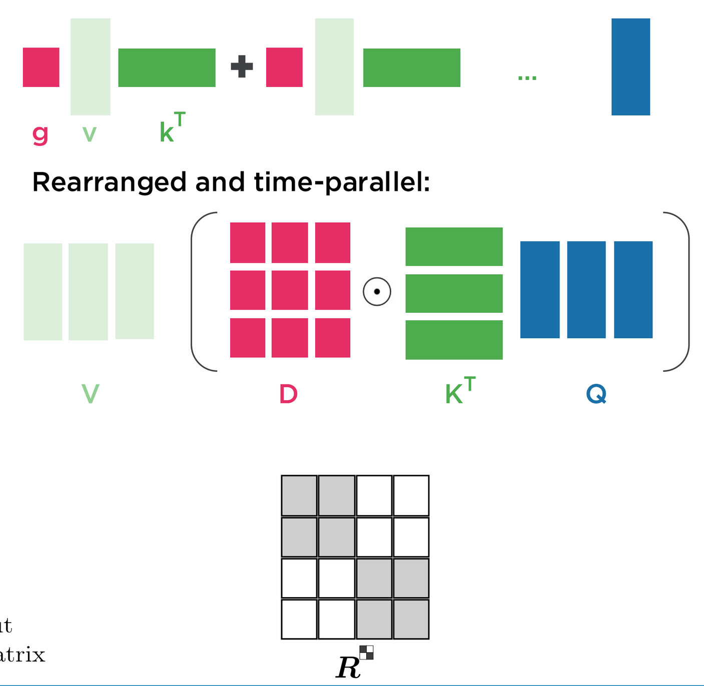
mLSTM

$$\{\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}, \tilde{\mathrm{i}}, \tilde{\mathrm{f}}, \tilde{\mathbf{o}}\}_t = \boldsymbol{W}_{\{q,k,v,\mathrm{i,f,o}\}} \boldsymbol{x}_t +$$

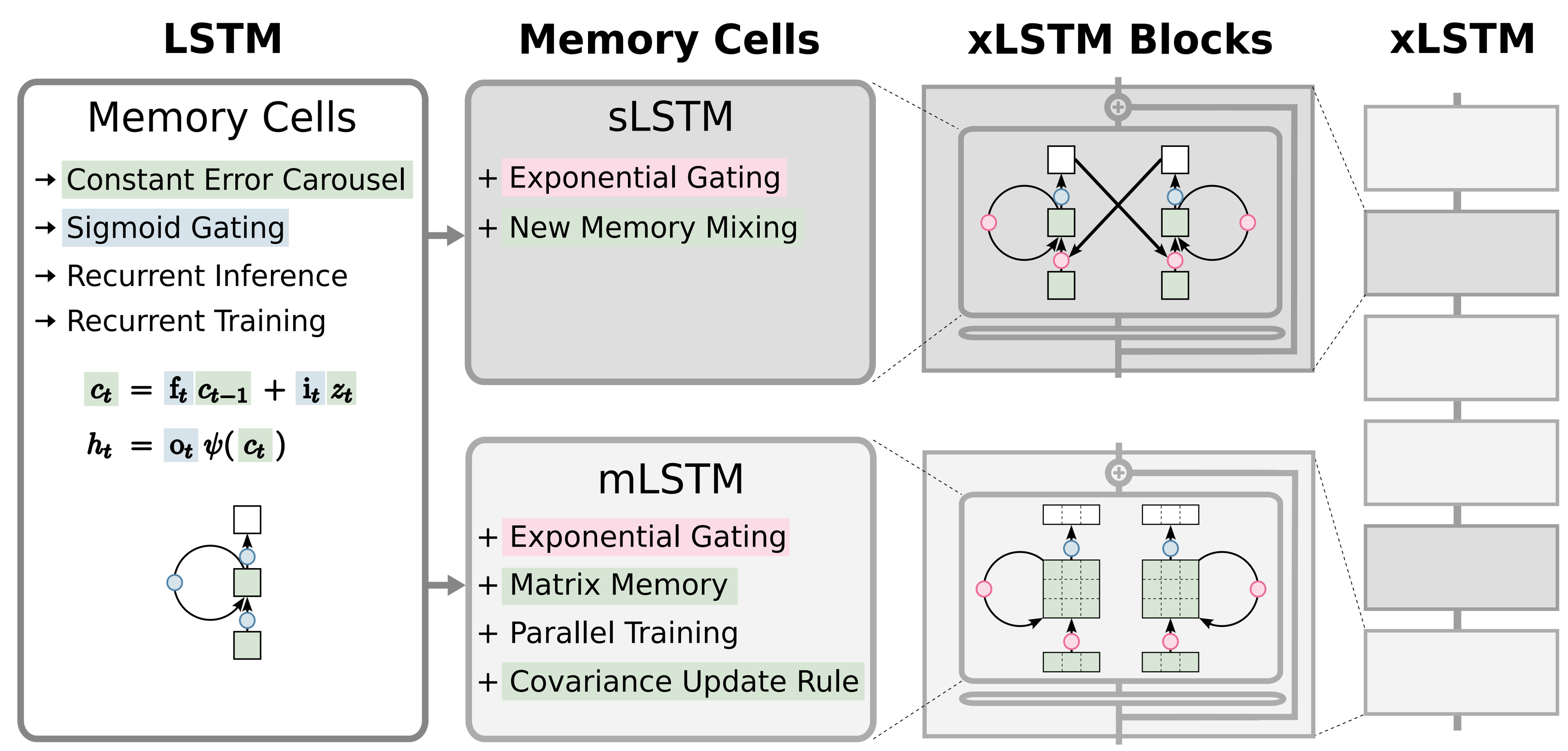
sLSTM

$$\{\tilde{\mathbf{i}}, \tilde{\mathbf{f}}, \tilde{\mathbf{z}}, \tilde{\mathbf{o}}\}_t = \boldsymbol{W}_{\{\mathrm{i,f,z,o}\}} \boldsymbol{x}_t + \boldsymbol{R}_{\{\mathrm{i,f,z,o}\}} \boldsymbol{h}_{t-1}$$

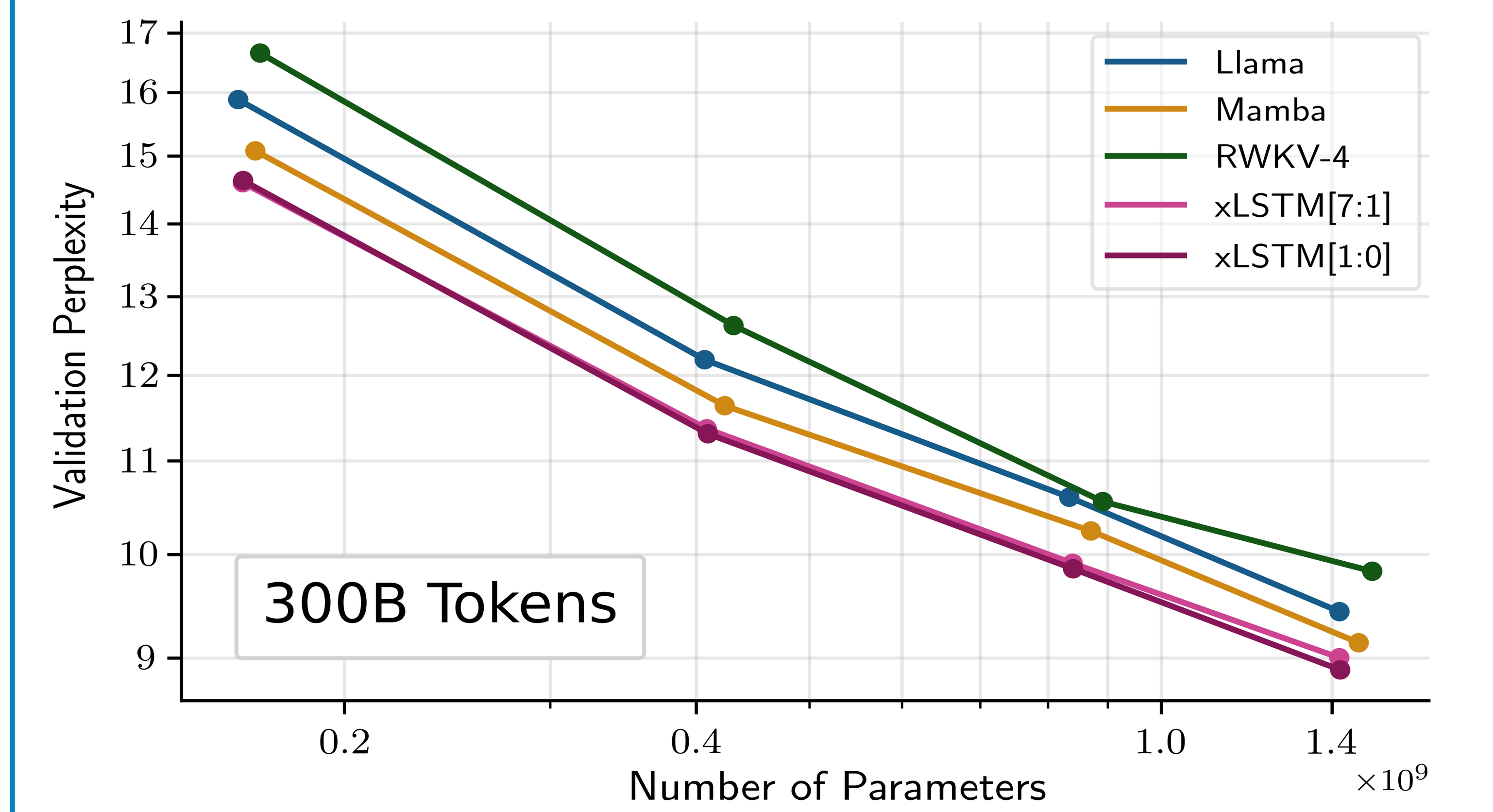
$\tilde{\mathrm{i}}, \tilde{\mathrm{f}}, \tilde{\mathbf{o}} \in \mathbb{R}^d$ or $\mathbb{R}$ input / forget / output gate preactivation
$\tilde{\boldsymbol{z}}_t, \boldsymbol{q}_t, \boldsymbol{k}_t, \boldsymbol{v}_t \in \mathbb{R}^d$ cell input, query, key, value
$\boldsymbol{x}_t \in \mathbb{R}^d$ input $\quad t \in \{1..T\}$ time
$\boldsymbol{W}_{\{i,f,z,o,q,k,v\}} \in \mathbb{R}^{d \times d}$ weight matrix
$\boldsymbol{c}_t \in \mathbb{R}^d, \boldsymbol{C}_t \in \mathbb{R}^{d \times d}$ cell state
$\boldsymbol{n}_t \in \mathbb{R}^d$ normalizer state
$\boldsymbol{h}_t \in \mathbb{R}^d$ hidden state / output
$\boldsymbol{R}_{\{i,f,z,o\}} \in \mathbb{R}^{d \times d}$ recurrent matrix



OVERVIEW



SCALING



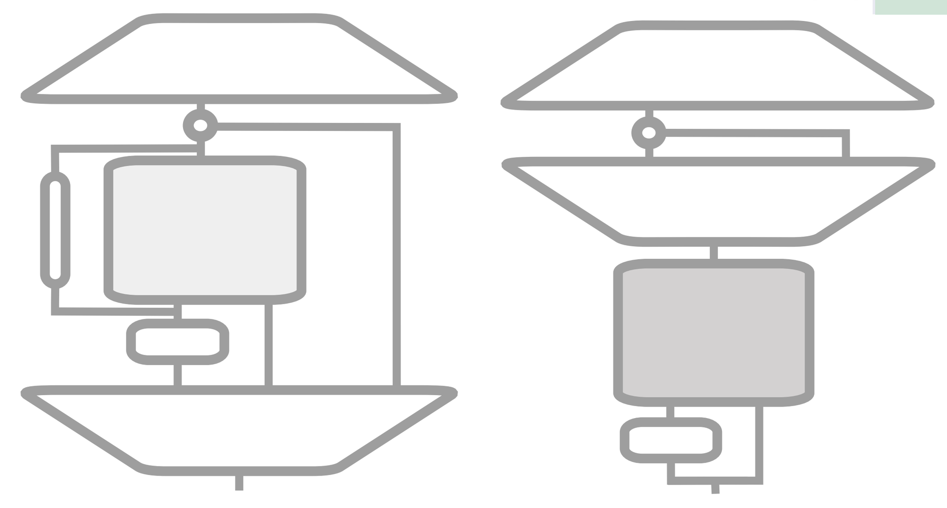
- Perplexity after training on 300B tokens of SlimPajama
- xLSTM scales similar to competitors with more parameters

SEQUENCE AND LANGUAGE MODELING

- xLSTM with sLSTM (memory mixing) can solve formal language tasks
- keeps the state tracking capabilities of LSTM

	Deterministic Context Free		Regular			
	Mod Arithmetic (w Brackets)	Solve Equation	Cycle Nav	Even Pairs	Mod Arithmetic (w/o Brackets)	Parity
Llama	0.02 ± 0.0	0.02 ± 0.0	0.04 ± 0.01	1.0 ± 0.0	0.03 ± 0.0	0.03 ± 0.0
Mamba	0.04 ± 0.01	0.05 ± 0.02	0.86 ± 0.04	1.0 ± 0.0	0.05 ± 0.02	0.13 ± 0.0
RWKV-6	0.09 ± 0.01	0.09 ± 0.02	0.31 ± 0.14	1.0 ± 0.0	0.16 ± 0.0	0.22 ± 0.1
LSTM	0.72 ± 0.04	0.38 ± 0.05	0.93 ± 0.07	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
xLSTM[0:1]	0.57 ± 0.09	0.55 ± 0.09	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
xLSTM[1:1]	0.15 ± 0.06	0.24 ± 0.04	0.8 ± 0.03	1.0 ± 0.0	0.6 ± 0.4	1.0 ± 0.0

- xLSTM Model Structure:
  ResNet-like architecture of Pre-Layer-Norm blocks
  mLSTM (pre-up projection) block
  sLSTM (post-up projection) block
  Combination: xLSTM[a:b] (mLSTM/sLSTM ratio)

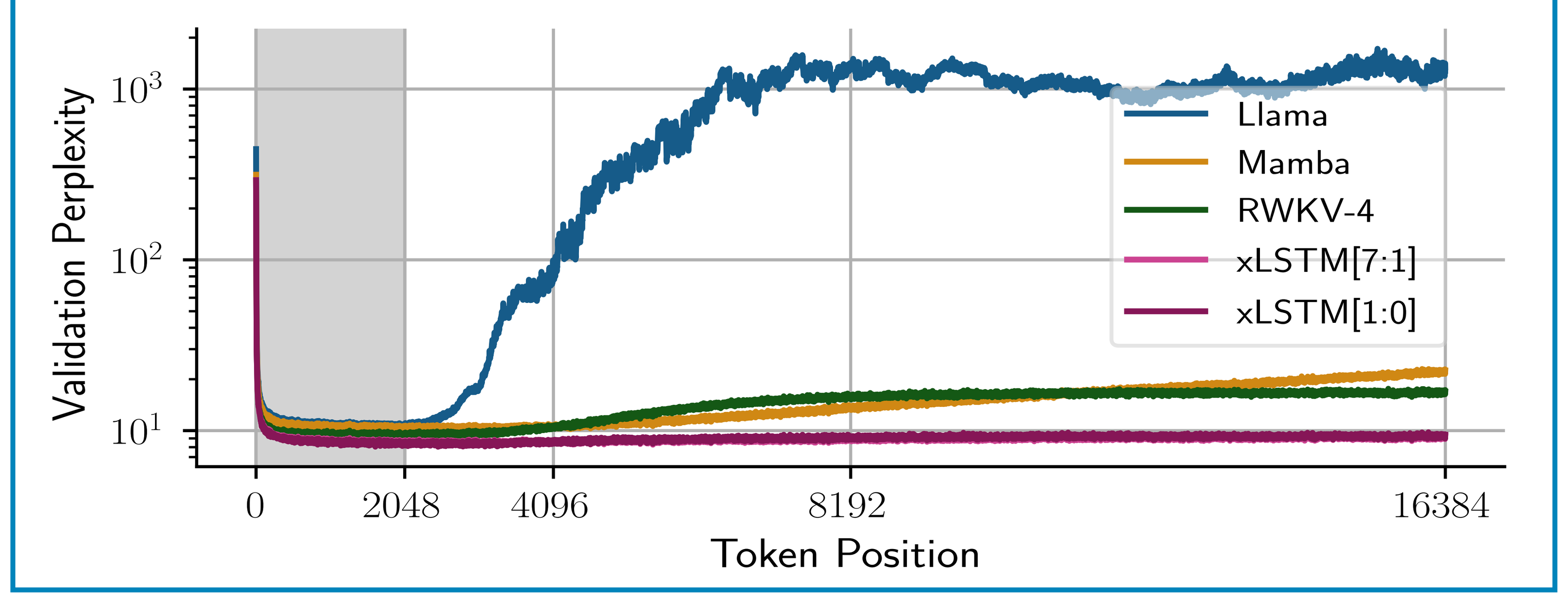


mLSTM block (pre-up projection) **sLSTM block** (post-up projection)

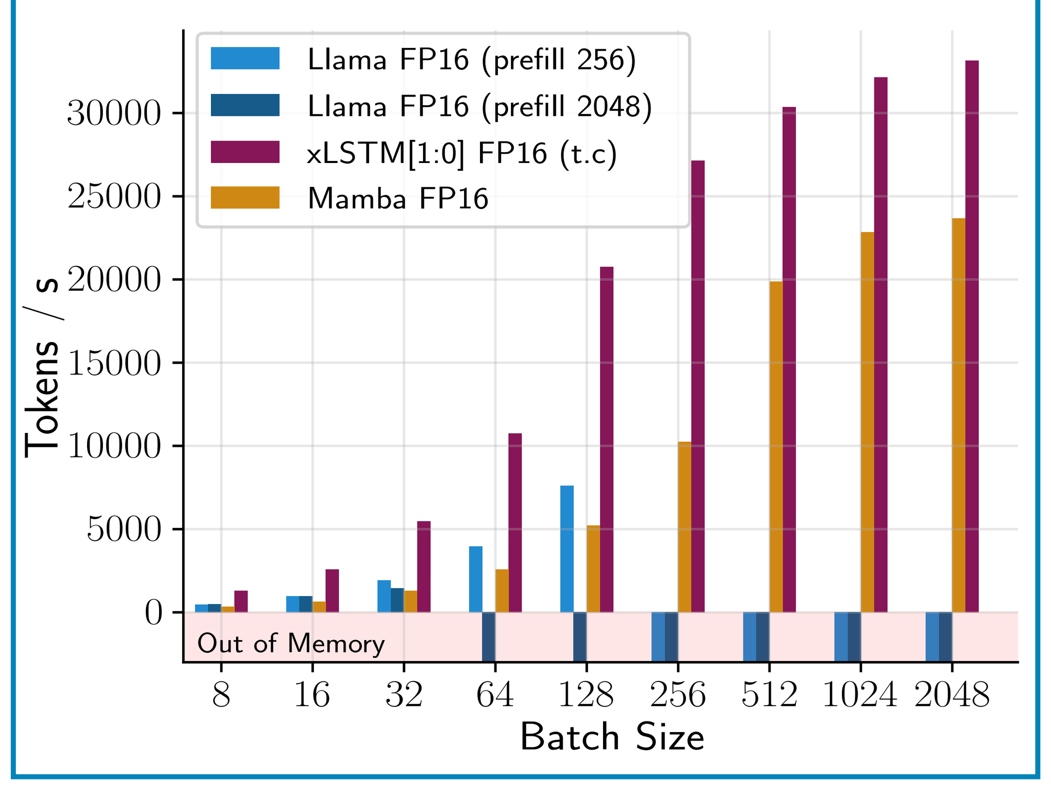
Model	#Params M	SlimPajama (15B) ppl ↓
GPT-3	356	14.26
Llama	407	14.25
H3	420	18.23
Mamba	423	13.70
Hyena	435	17.59
RWKV-4	430	15.62
RWKV-5	456	14.25
RWKV-6	442	15.03
RetNet	431	16.23
HGRN	411	17.59
GLA	412	16.15
HGRN2	411	14.32
xLSTM[1:0]	409	**13.43**
xLSTM[7:1]	408	13.48

- Competitive scores on Multi-Query Associative Recall (memory) and Long-Range Arena (long-range) tasks
- State-of-The Art Language Modeling Perplexity on SlimPajama (15B) at 350M parameter scale
- Downstream Performance on LMEval and PALOMA tasks matches PPL performance gap

LENGTH EXTRAPOLATION



GENERATION SPEED



Paper:
arXiv:2405.04517

Code:
github.com/NX-AI/xlstm