L^2M: Mutual Information Scaling Law for Long-Context Language Modeling

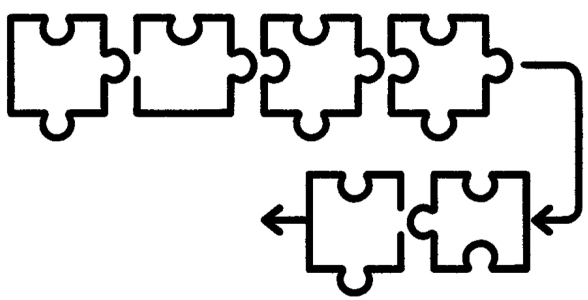
Zhuo Chen[12]*, Oriol Mayné i Comas[23], Zhuotao Jin[24], Di Luo[1245], Marin Soljačić[12]

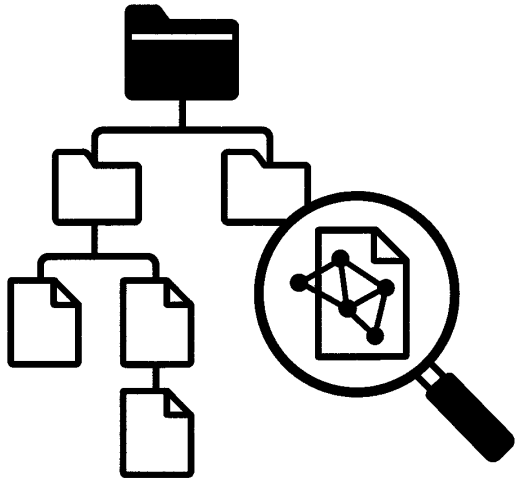
[1]The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
[2]Massachusetts Institute of Technology
[3]Polytechnic University of Catalonia
[4]Harvard University
[5]University of California, Los Angeles

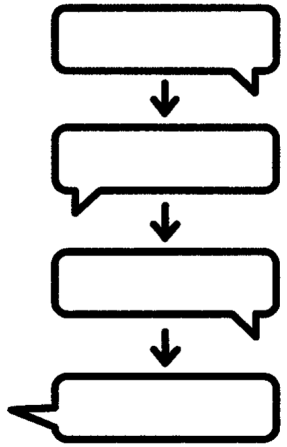
* chenzhuo@mit.edu
https://arxiv.org/abs/2503.04725

Long-context Language Modeling is Crucial

- **Reasoning through long chains of logic**
- **Maintaining consistency in long writing**
- **Understanding large codebases**
- **Keeping track of an extended conversation**





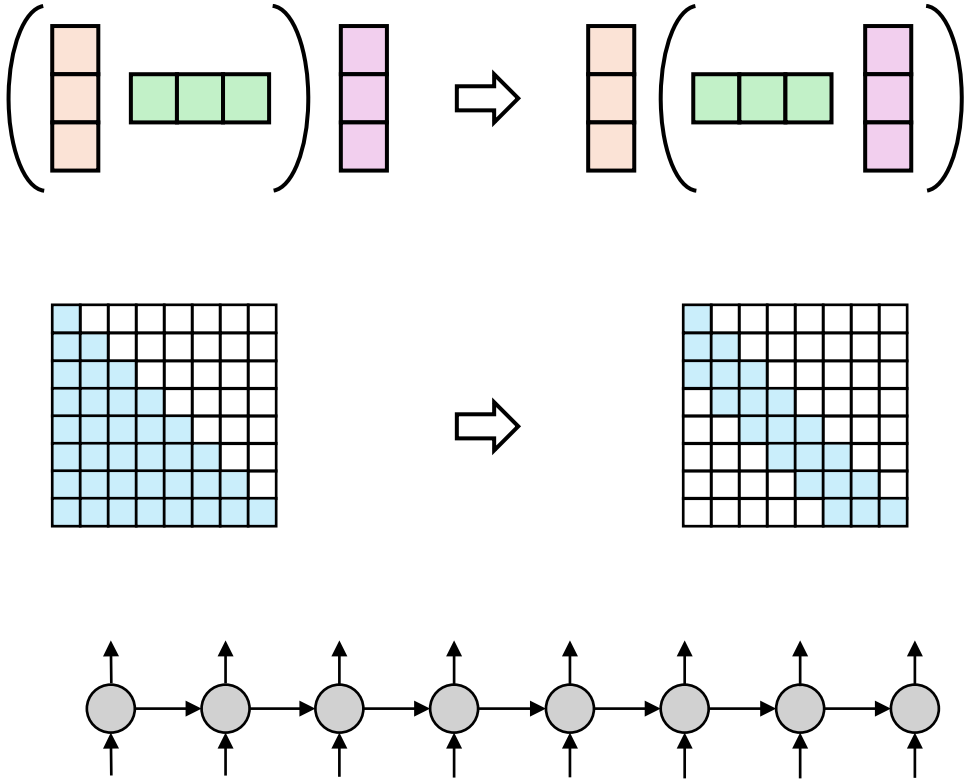


Limited Understanding of Trade-off Between Cost and Performance

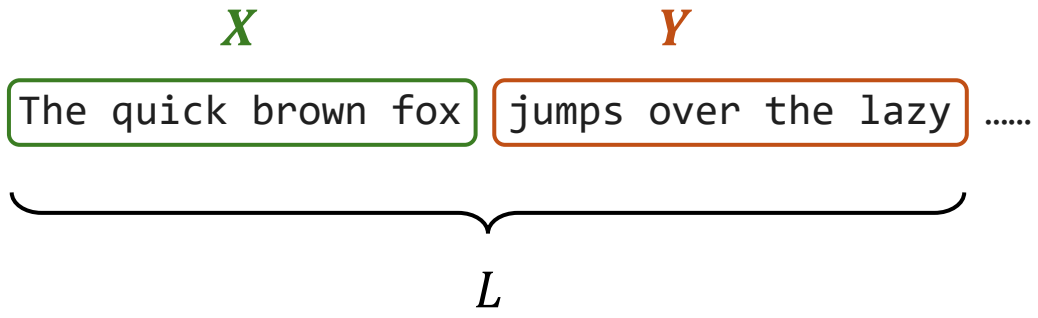
Transformers have quadratic computational cost

- Linear attention
- Sparse attention
- State space model
- Compression
- OCR
- ……

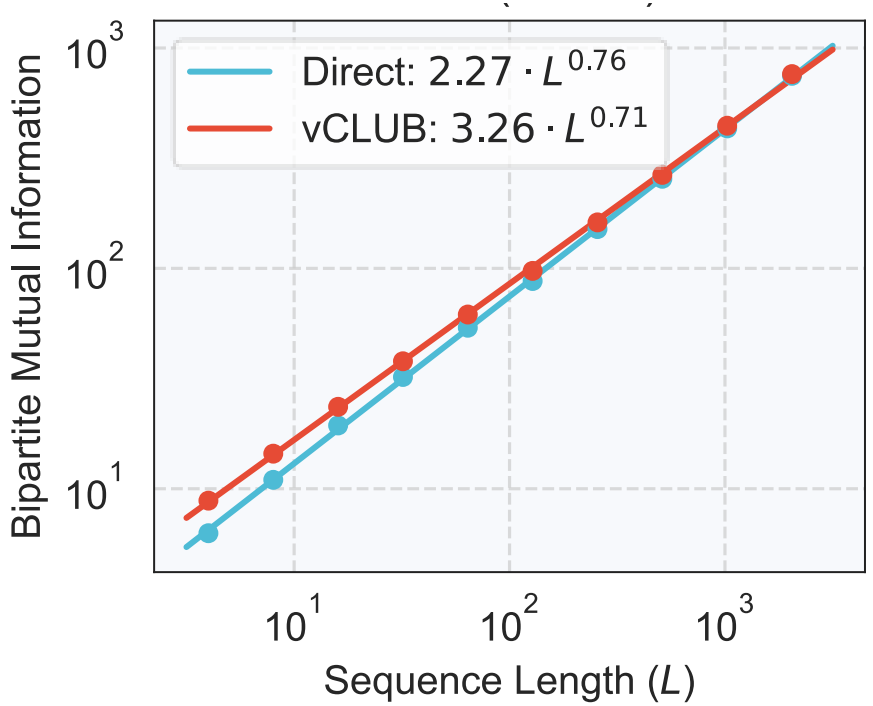
Lower cost without hurting performance?



Universal Framework Based on Bipartite Mutual Information

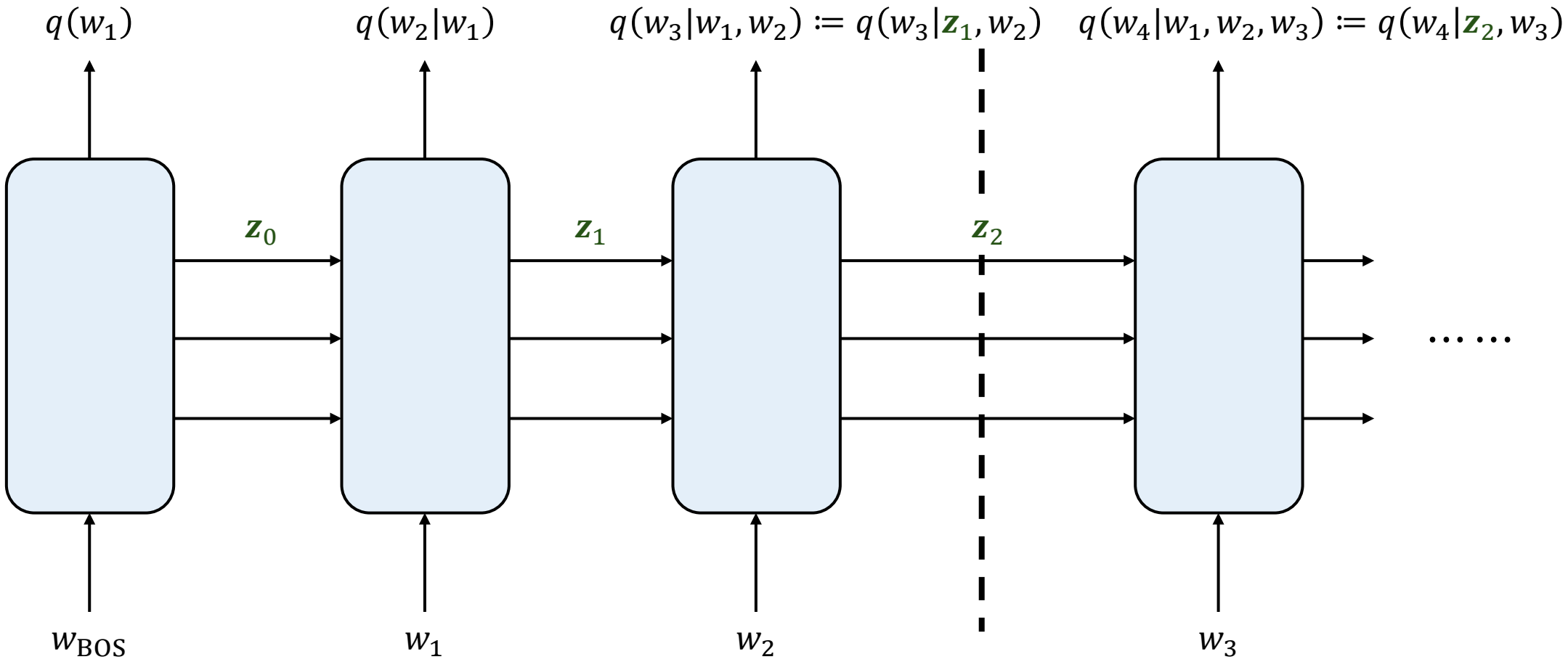


Bipartite mutual information $I(X;Y) \sim L^{\beta}$



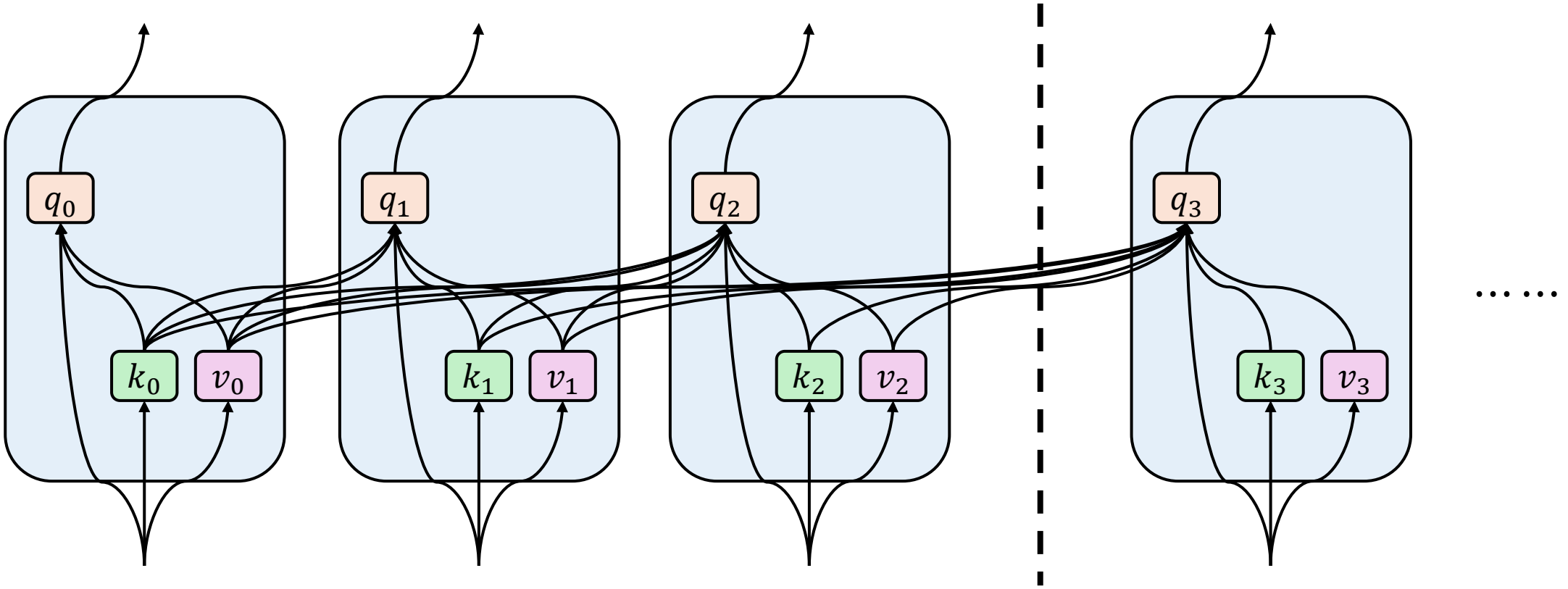
Measures the amount of information that must be retained from X to correctly predict $p(Y|X)$

History State in Large Language Models



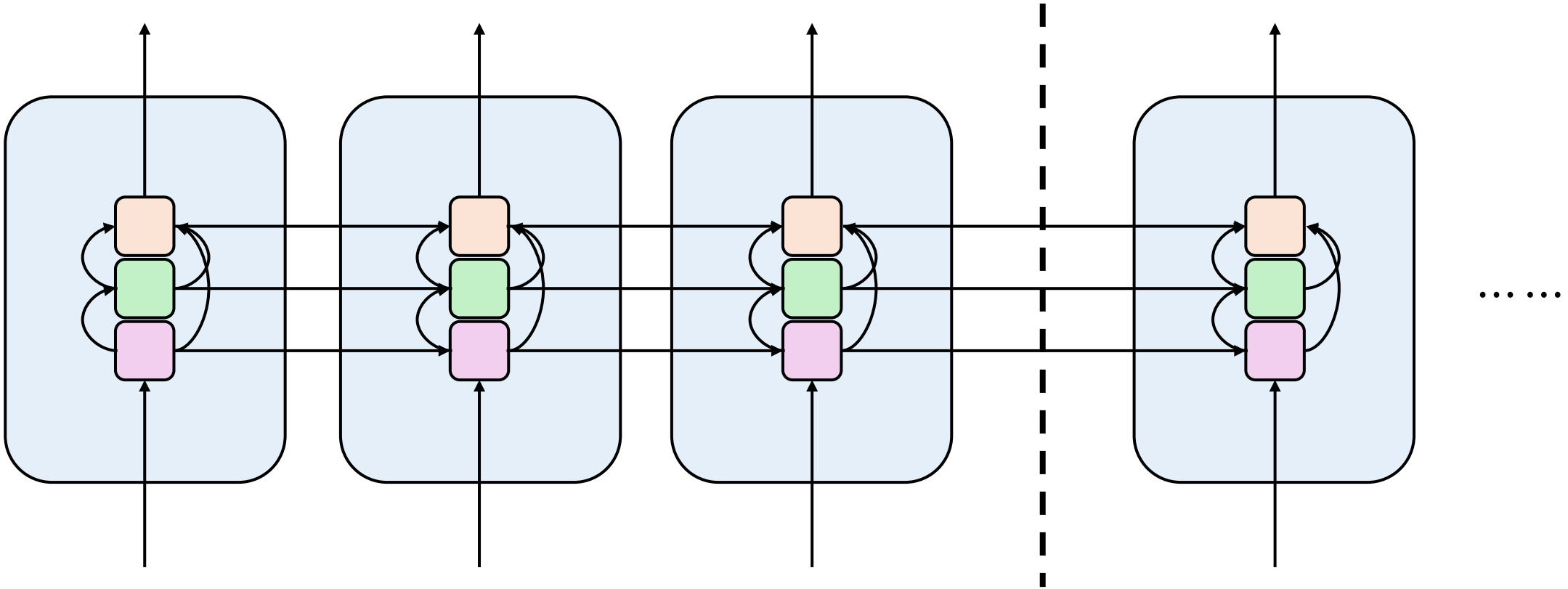
History state = the latent variables responsible for storing past information

History State in Transformers



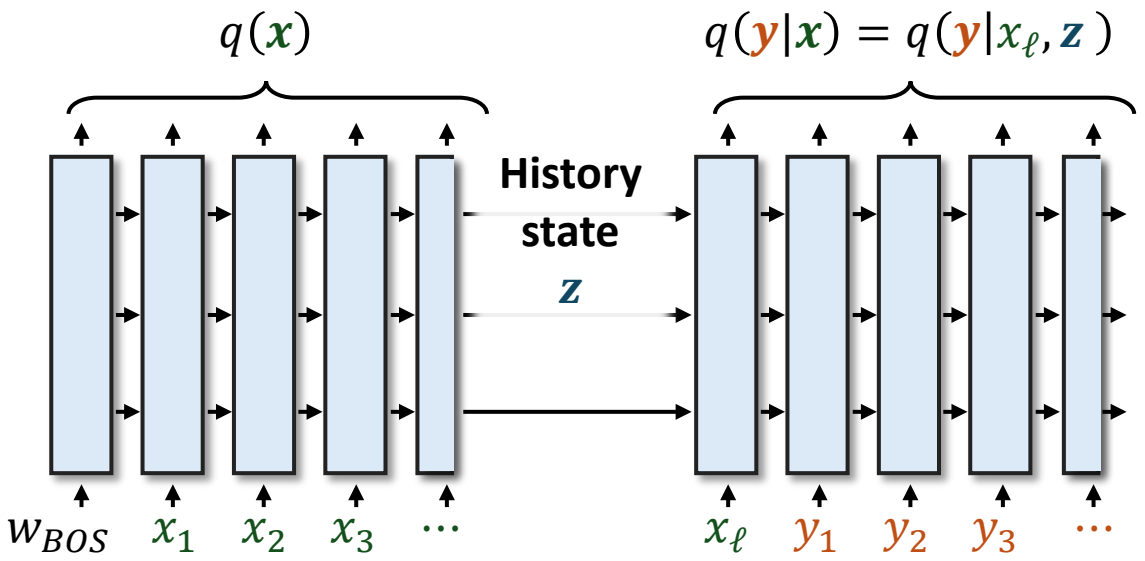
History state = all key-value pairs

History State in State Space Models

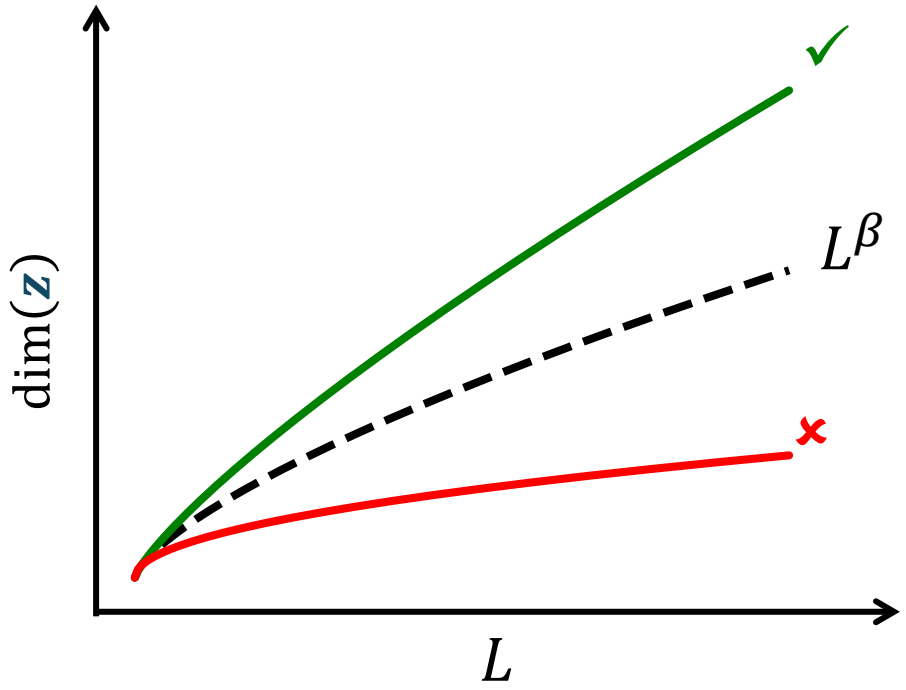


History state = recurrent state

Necessary Scaling of History State

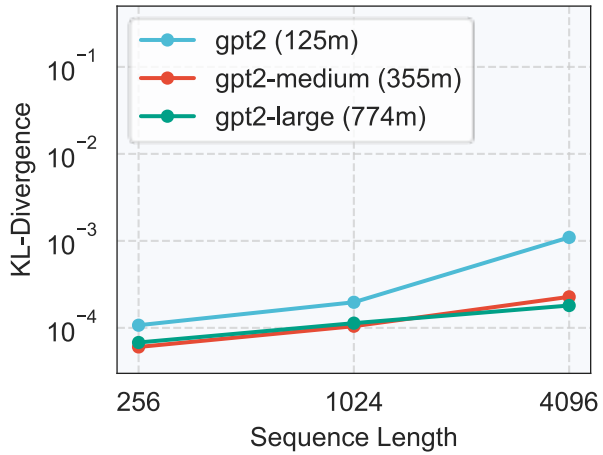


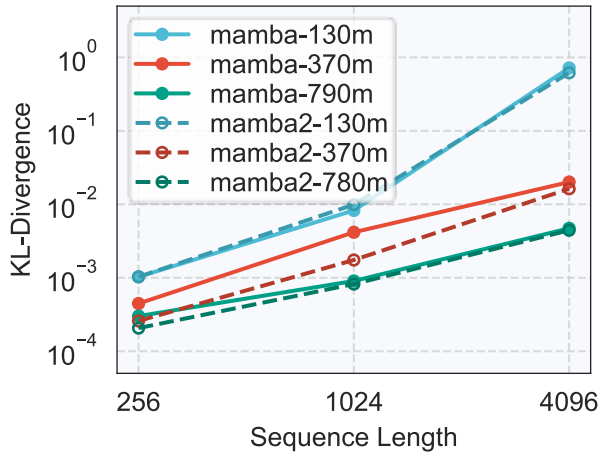
$$\max I^q(X;Y) \sim \dim(z)$$

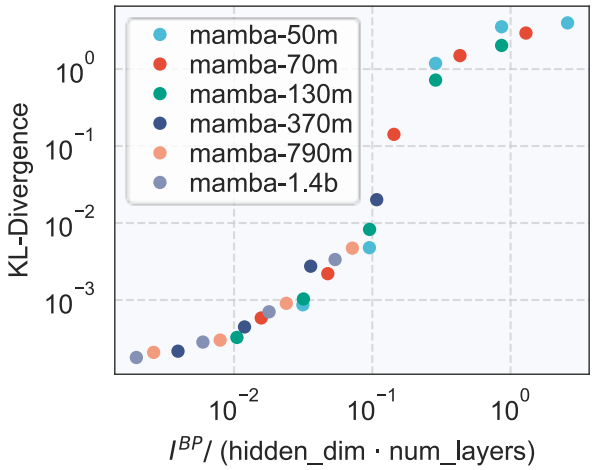


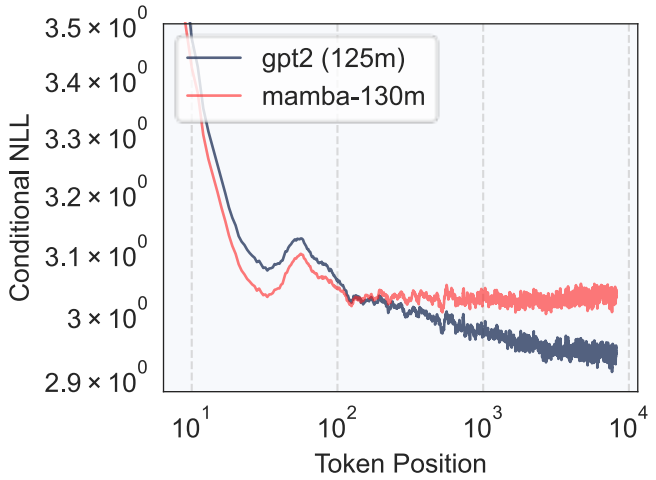
An LLM architecture is effective for long context modeling only if $\dim(z) \gtrsim L^\beta$

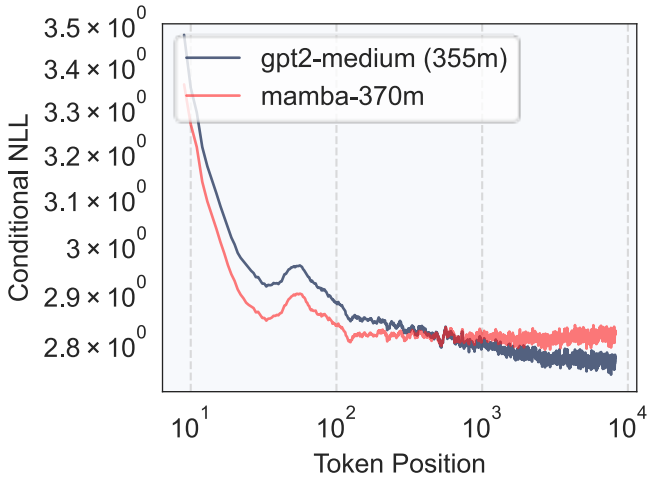
Experimental Confirmation of Our Theory

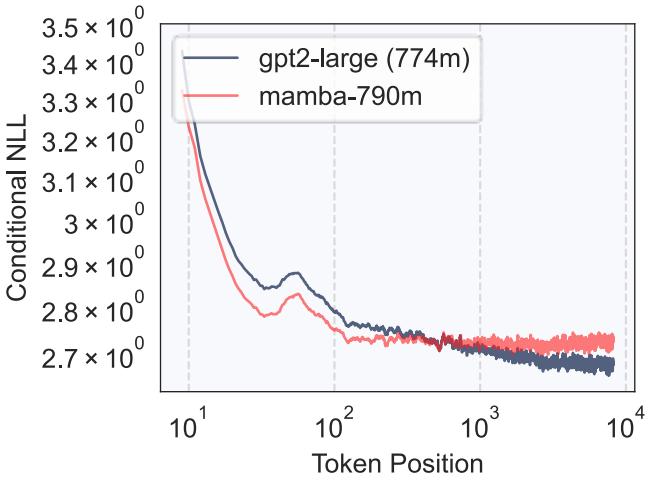










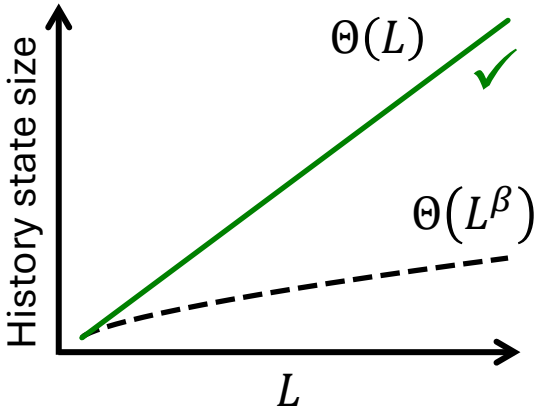


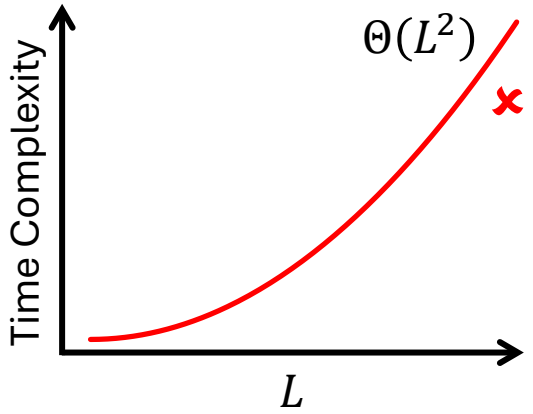
Neither Transformer nor State Space Model is Optimal

Transformers:

✓ History state size ~ $\Theta(L)$

× Time complexity ~ $\Theta(L^2)$

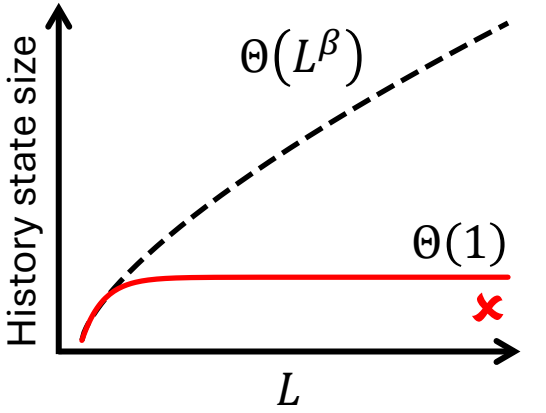


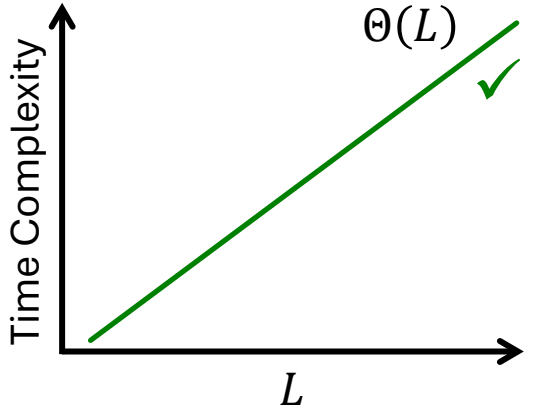


State space models:

× History state size ~ $\Theta(1)$

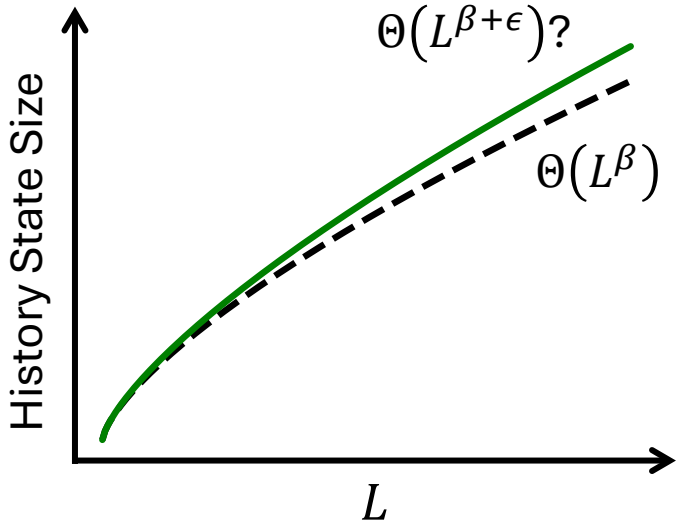
✓ Time complexity ~ $\Theta(L)$



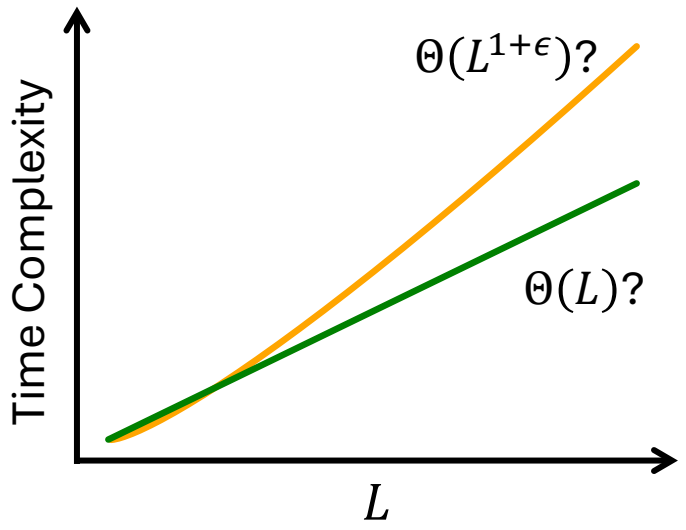


Optimal Level of Sparsity?

We don’t need linearly scaling history size



and we want low time complexity



Future direction: to engineer the optimal MI-capable sparsity