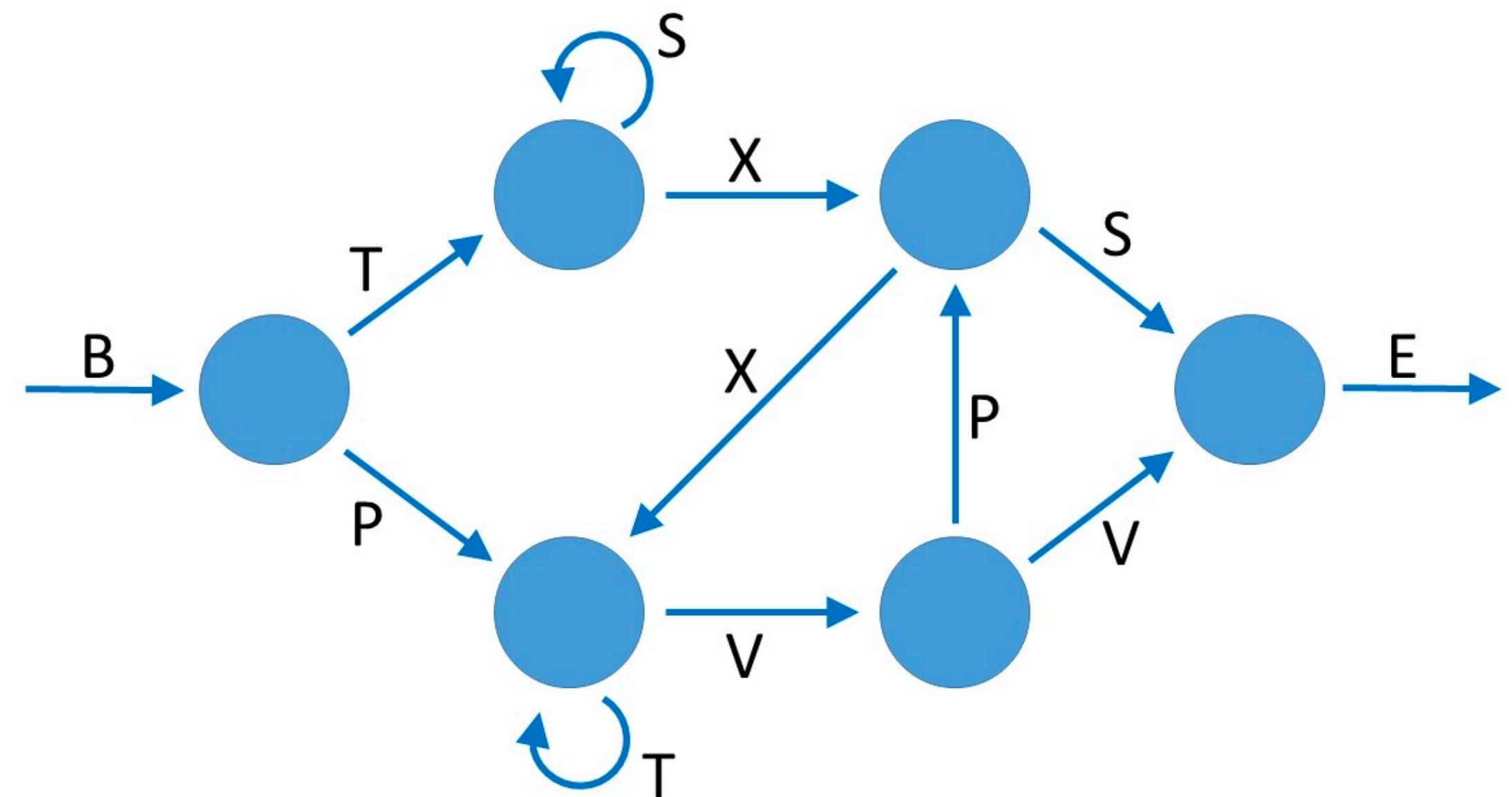


In Defense of Small Language Models

Sensory Prediction: Engineered and Evolved

Santa Fe Institute

12 July 2023



David Pfau



Natural Language Prediction

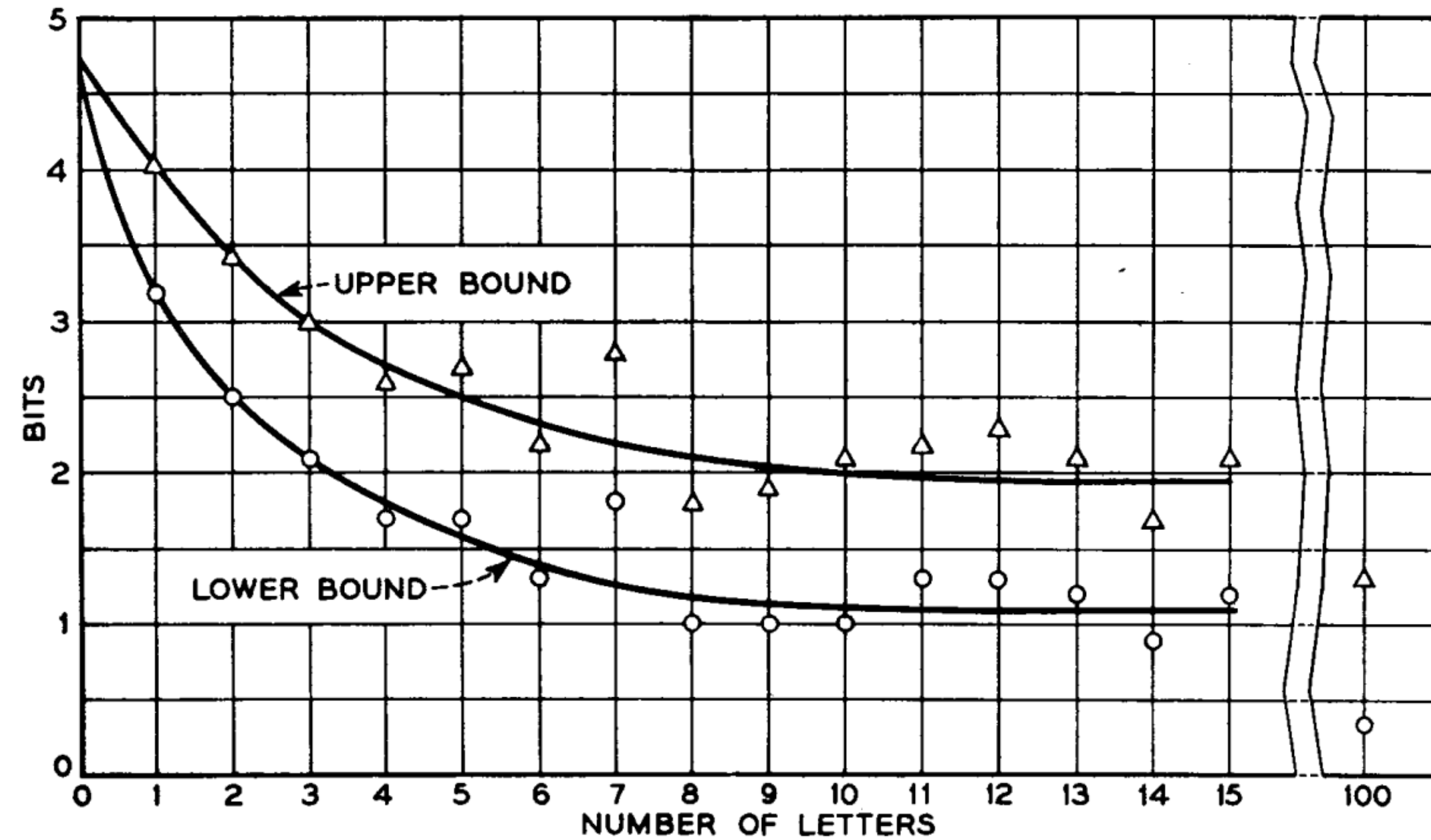
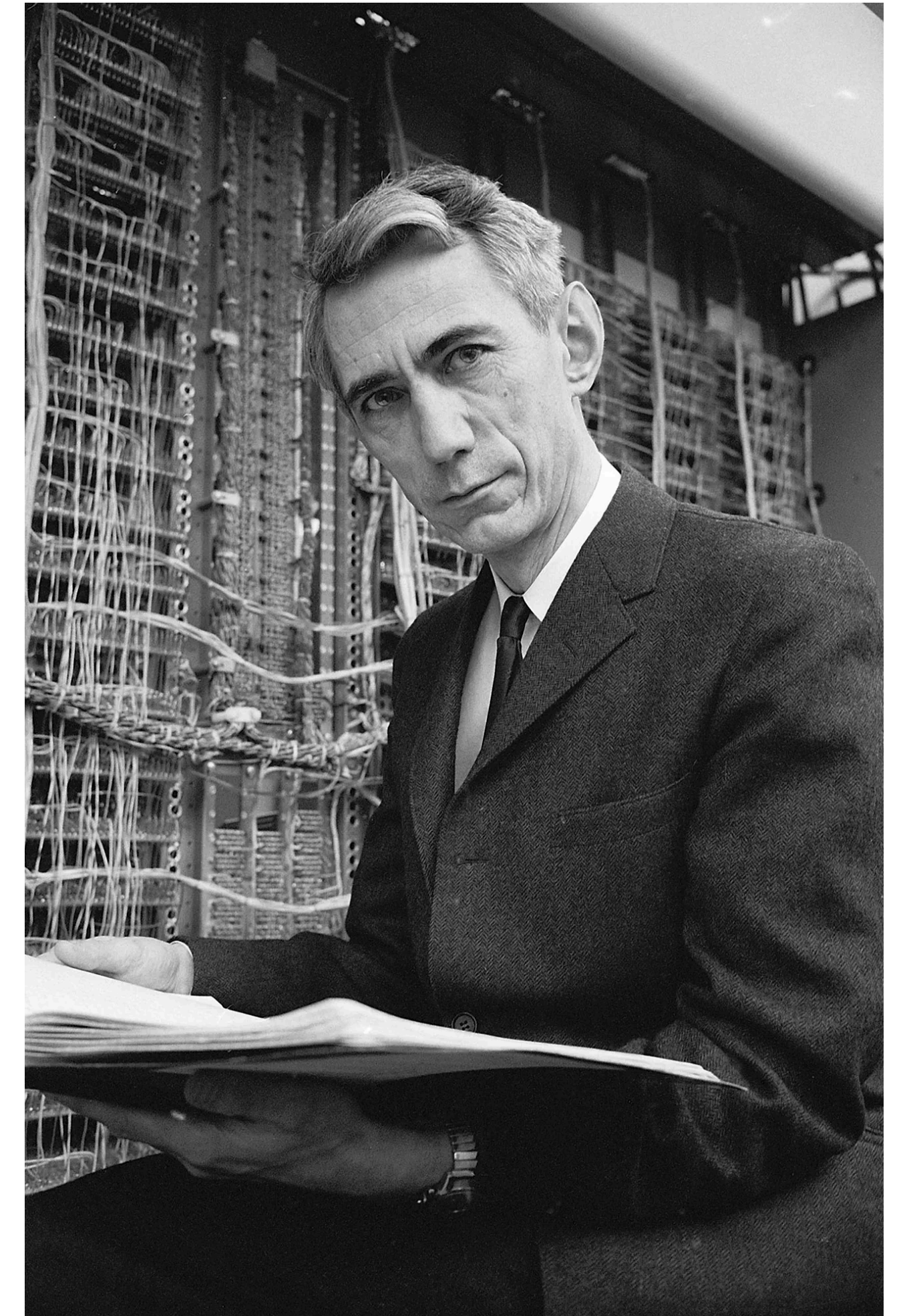


Fig. 4—Upper and lower experimental bounds for the entropy of 27-letter English.

Entropy rate of English text is ~ 1.3 bits per character

C. E. Shannon, Bell Systems Technical Journal (1951)

T. M. Cover and R. C. King, IEEE Trans. Information Theory (1978)



Prediction as Intelligence

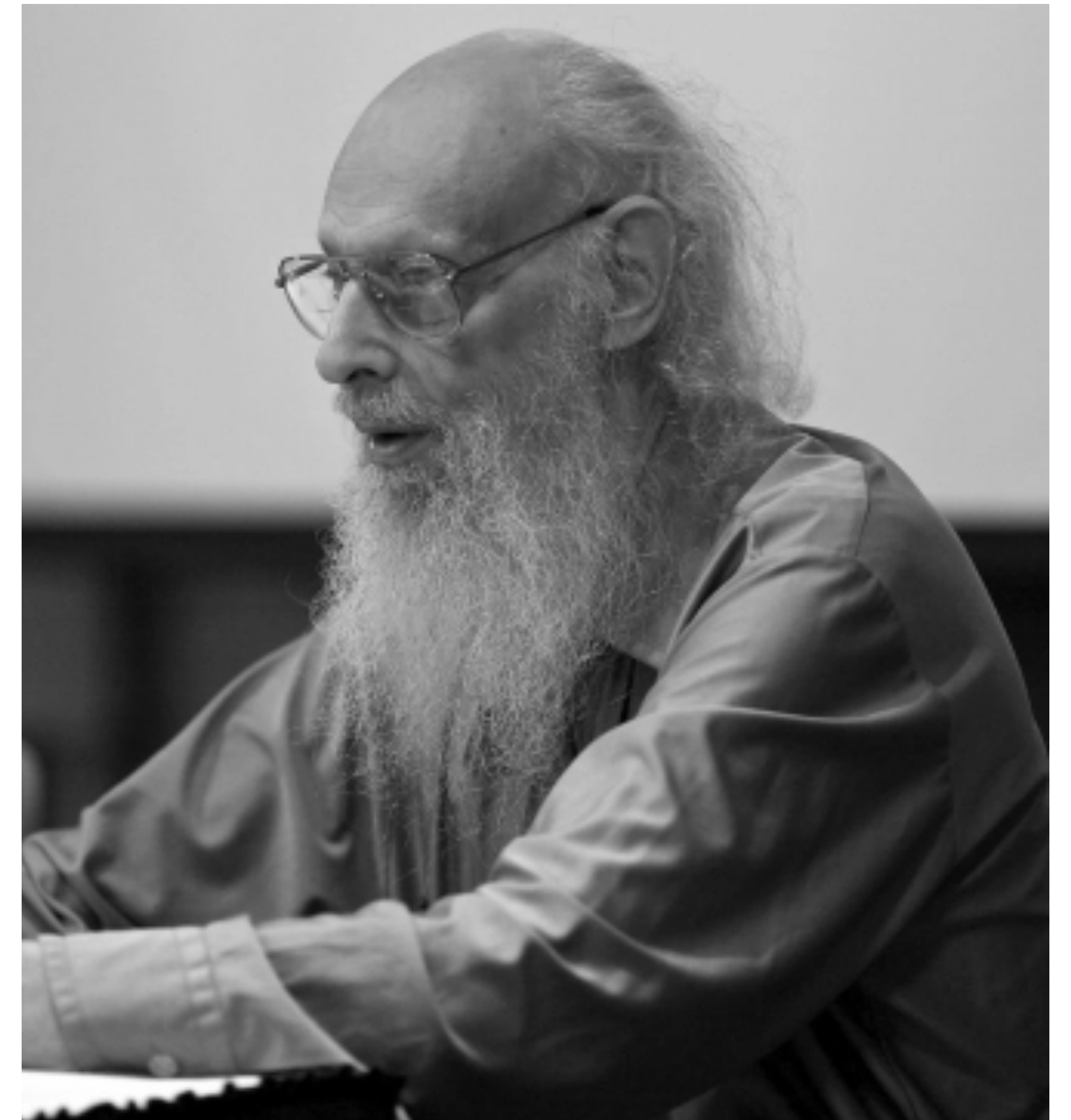
Solomonoff Induction

$$p(x_{N+1} | x_{1:N}) \propto p_{\theta}(x_{1:N+1}) \sum_{\theta \in \mathcal{M}} 2^{-\ell(\theta)}$$

Description length of **data** Description length of **model**

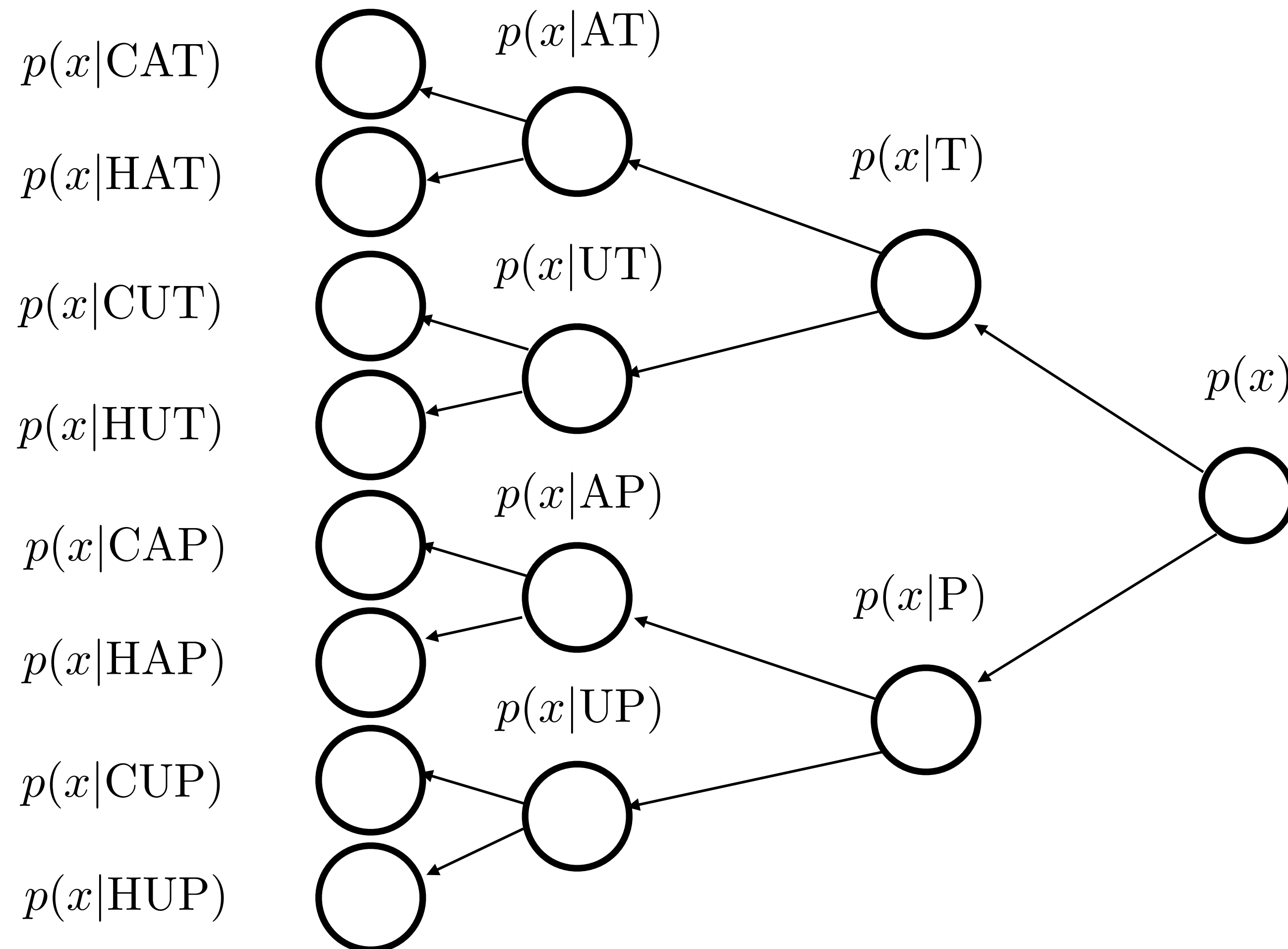
- Optimal induction is minimizing description length of data and model, model = program
- *Provably* incomputable!
- But builds a bridge between AI and info theory
- Inspirational to AIXI [M. Hutter (2000)], Wikipedia compression prize [M. Hutter (2006)]

R. J. Solomonoff, Information and Control (1964)



N-Gram Models

Language Modelling in the Before Times



$$p(x_{1:N}) = \prod_i p(x_i | x_{1:i-1})$$

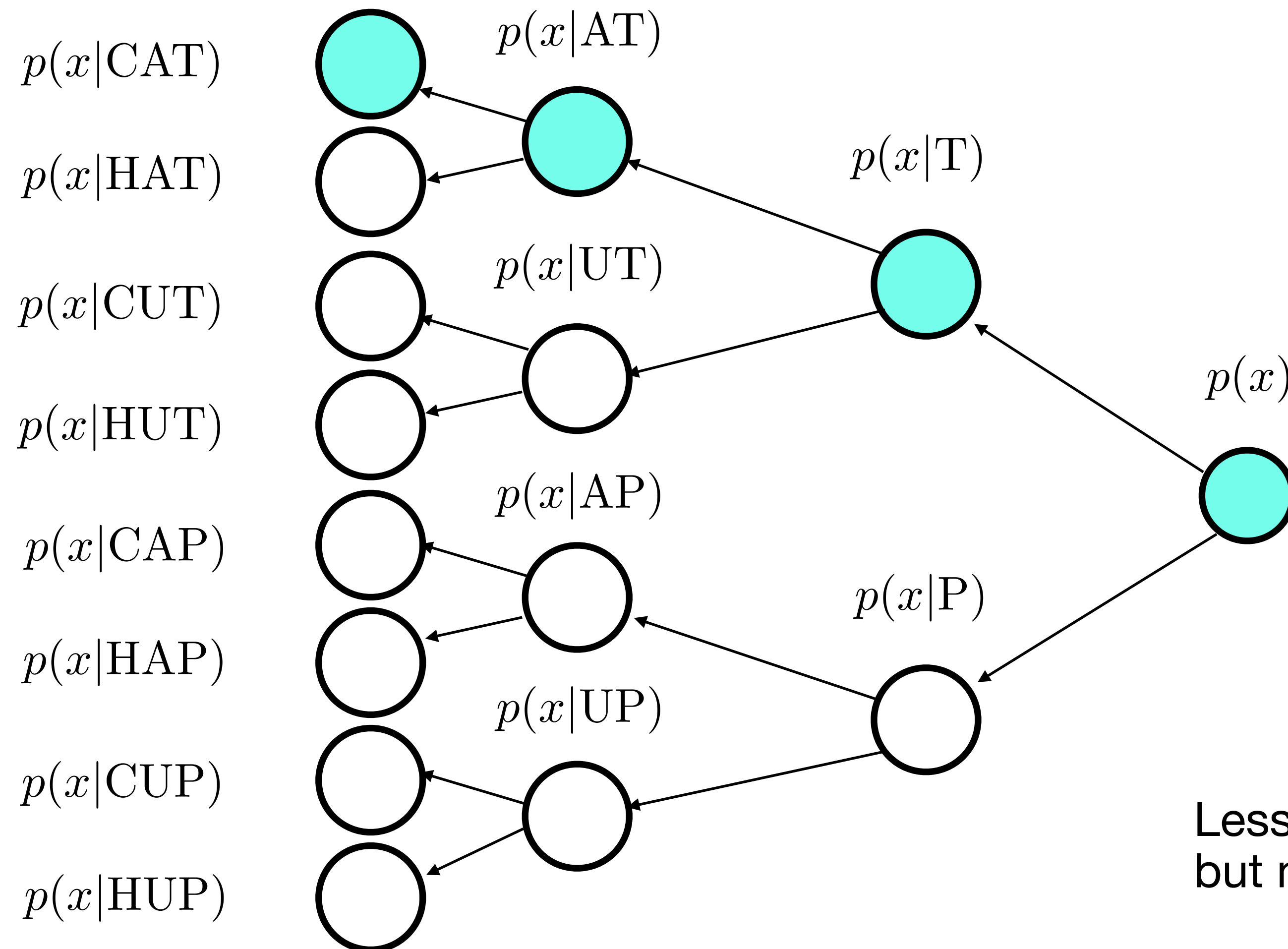
$$p(x_i | x_{1:i-1}) = p(x_i | x_{i-n:i-1})$$

$$\mathcal{S}(x_{-\infty:0}) = x_{-n:0}$$

WHEN_THE_CAT...

N-Gram Models

Language Modelling in the Before Times



$$p(x_{1:N}) = \prod_i p(x_i | x_{1:i-1})$$

$$p(x_i | x_{1:i-1}) = p(x_i | x_{i-n:i-1})$$

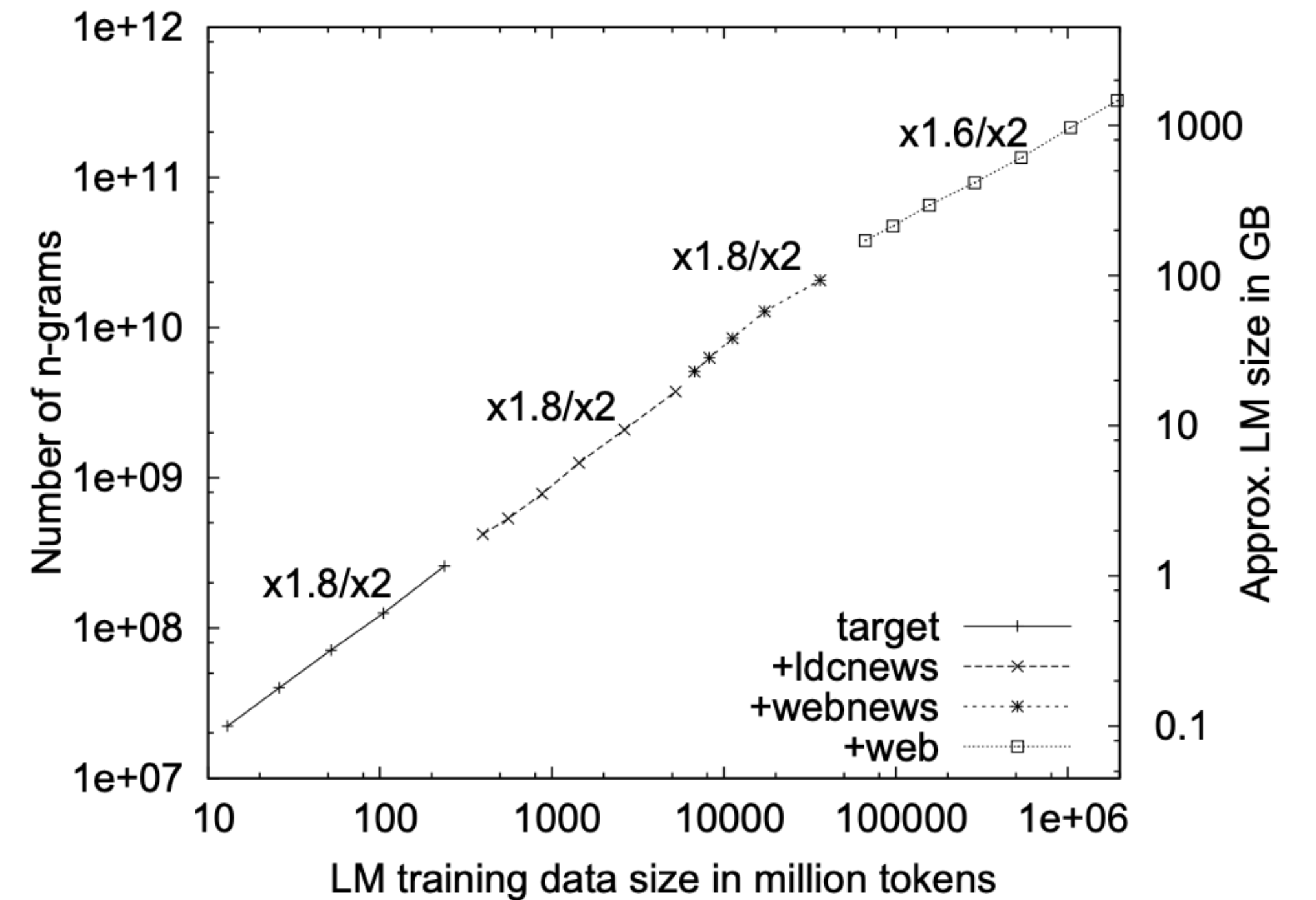
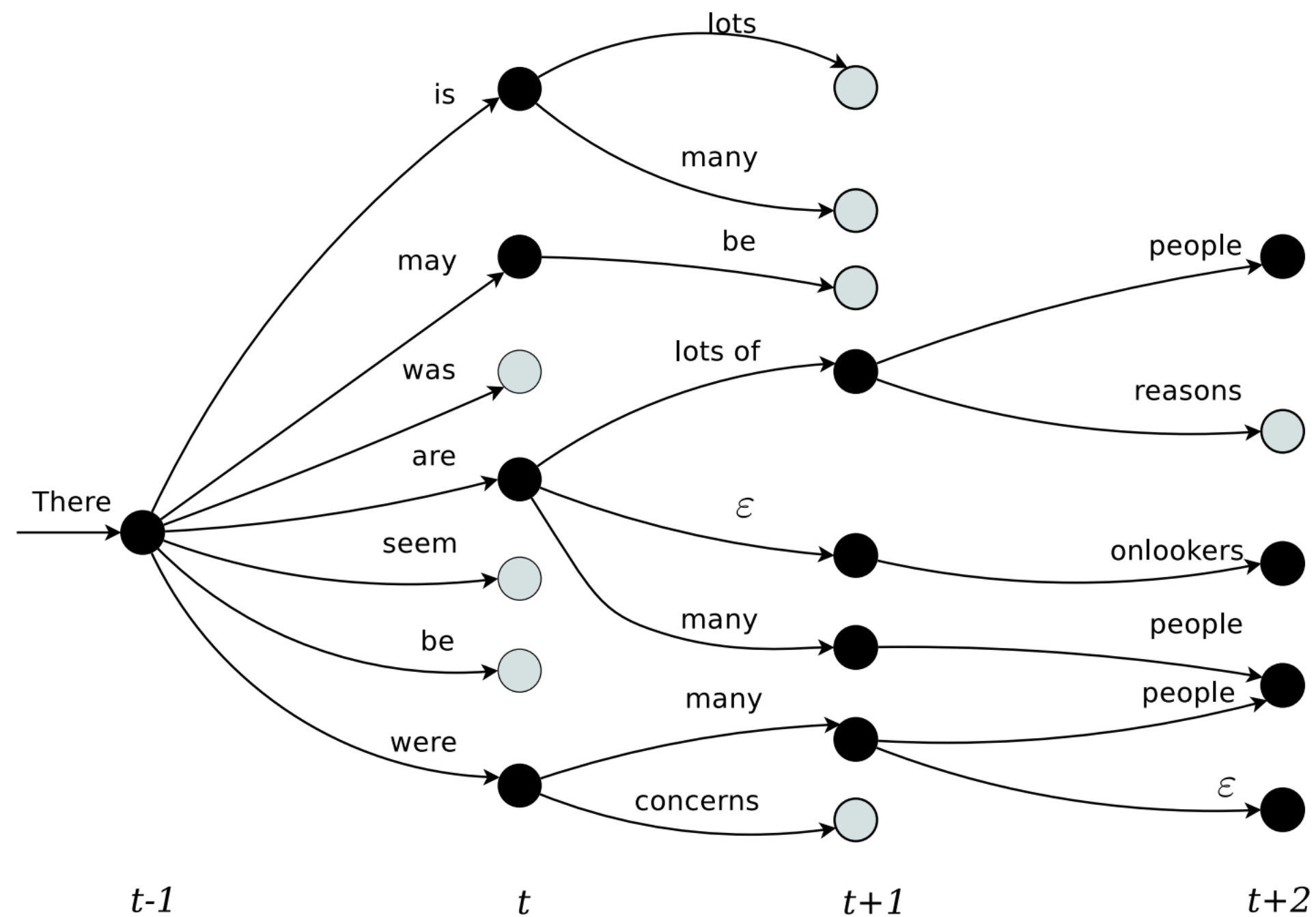
$$\mathcal{S}(x_{-\infty:0}) = x_{-n:0}$$

WHEN_THE_CAT...

Less expressive than Hidden Markov Models...
but more accurate at predicting natural language

N-Gram Models

Language Modelling in the Before Times



2 trillion tokens and 300 billion parameters - still large by today's standards!

ϵ -Machine

Computational Mechanics

$$\mathcal{I}[X_{-\infty:0}; X_{1:\infty}]$$

$$\mathcal{I}[\mathcal{S}(X_{-\infty:0}); X_{1:\infty}] = \mathcal{I}[X_{-\infty:0}; X_{1:\infty}]$$

$$\forall \mathcal{S} \exists f_{\mathcal{S} \rightarrow \mathcal{S}^*} \text{ s.t. } f_{\mathcal{S} \rightarrow \mathcal{S}^*} \circ \mathcal{S} = \mathcal{S}^*$$

$$\mathcal{S}^*(x_{-\infty:1}) = \delta(x_1, \mathcal{S}^*(x_{-\infty:0}))$$

- Resulting model class is *probabilistic state machine*
- Can do frequentist estimation - CSSR [Shalizi and Klinkner 2004]
- Can we do *Bayesian* estimation of ϵ -machines?

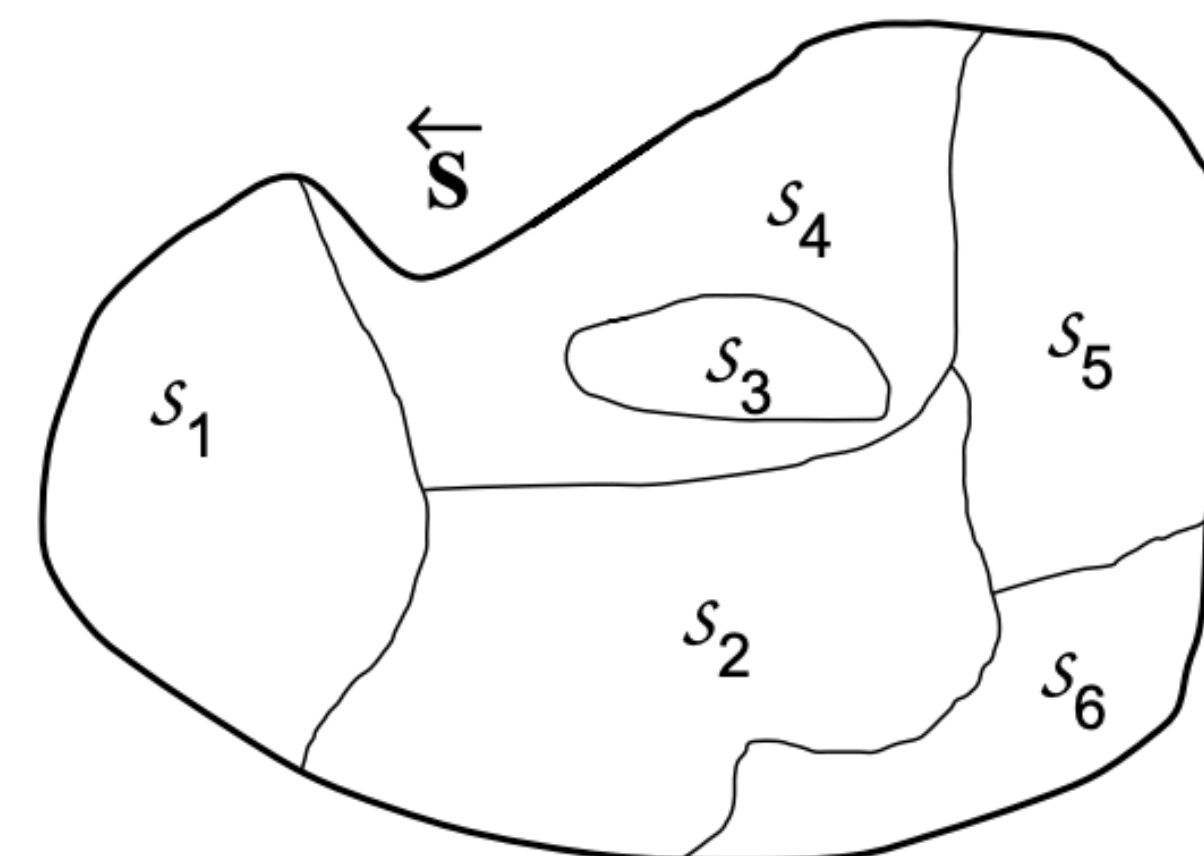
C. R. Shalizi and J. P. Crutchfield, Journal of Statistical Physics (2001)

Predictive Information

Sufficient statistic of the past

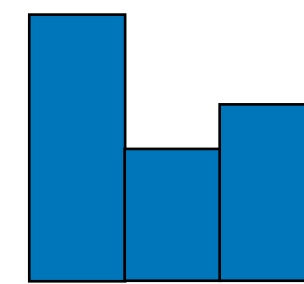
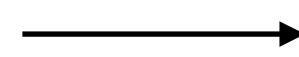
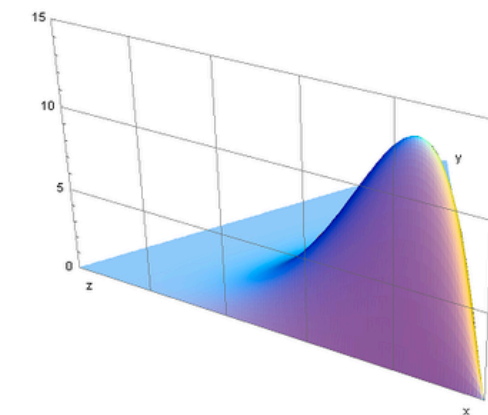
Minimal sufficient statistic of the past

Deterministic transition function



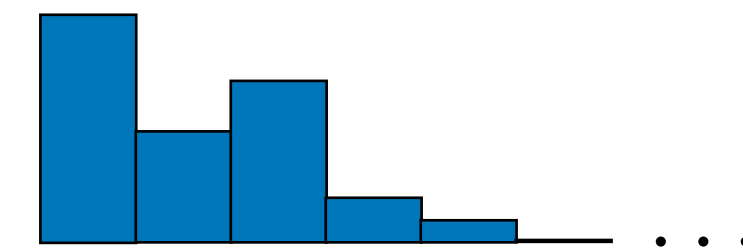
Nonparametric Bayesian Inference

$$\vec{\pi} \sim \text{Dir}(\vec{\alpha})$$



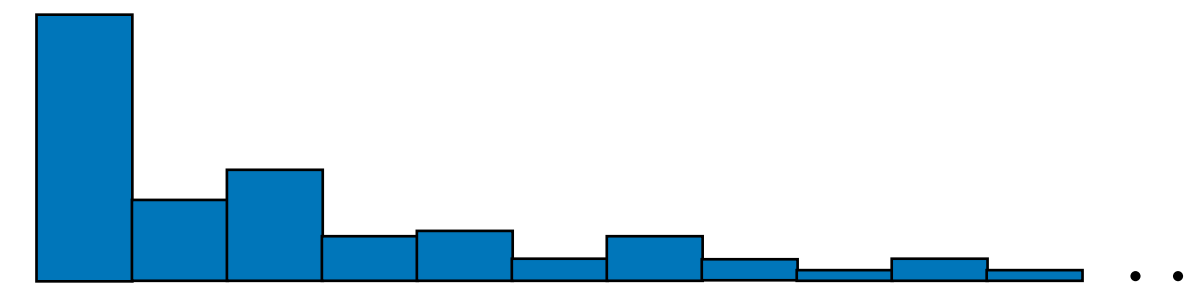
Dirichlet distribution

$$\vec{\pi} \sim \mathcal{DP}(\vec{\alpha}, \mu)$$



Dirichlet process

$$\vec{\pi} \sim \mathcal{PY}(\vec{\alpha}, d, \mu)$$



Pitman-Yor process

$\mathcal{O}(1)$

$\mathcal{O}(\log(N))$

$\mathcal{O}(N^\alpha)$

$\mathcal{O}(N)$



Parametric

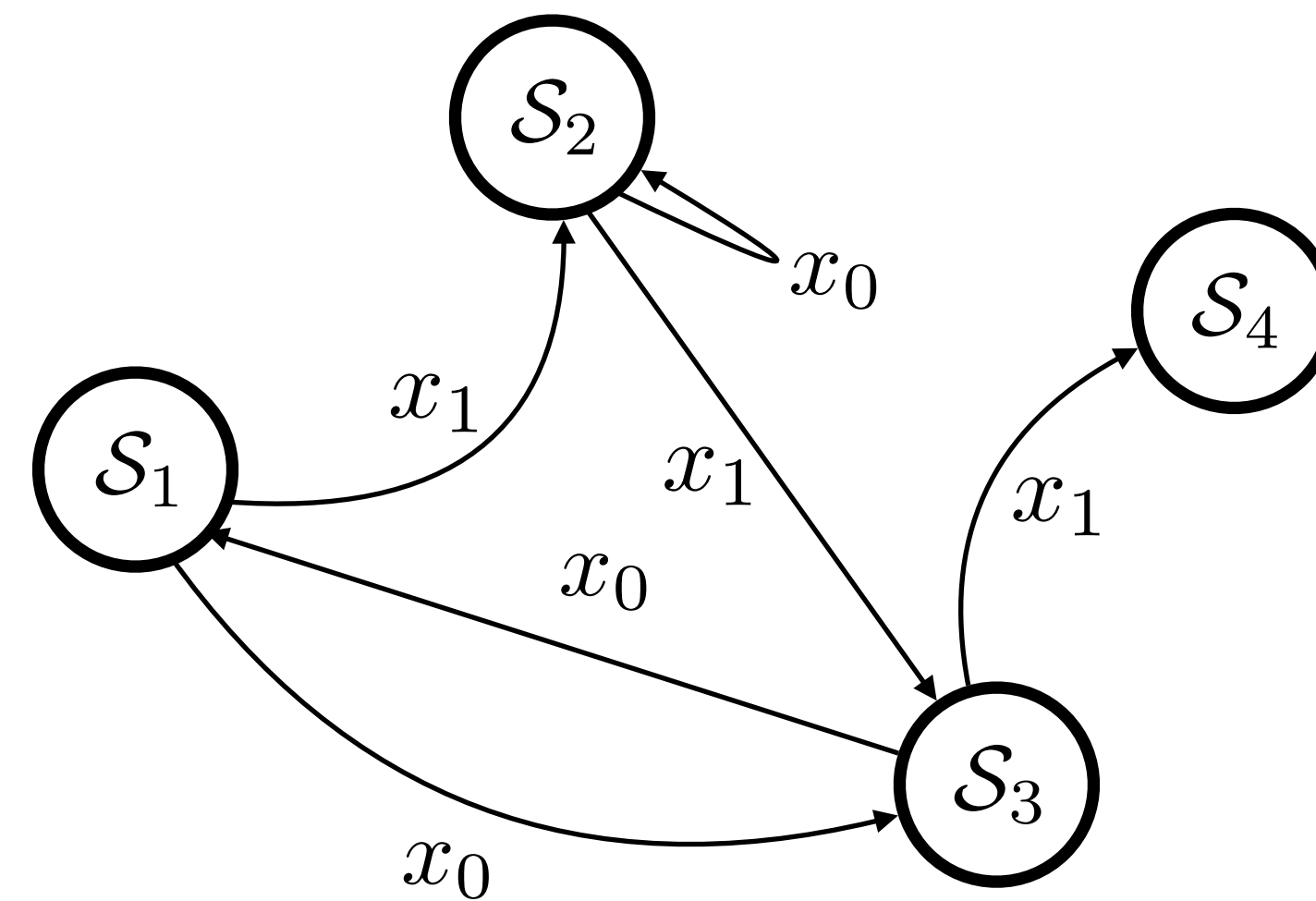
Bayesian nonparametric

“Classic” nonparametric

- Define prior probability over *infinite dimensional* objects
- Intermediate between parametric and nonparametric statistics

Probabilistic Deterministic Infinite Automata

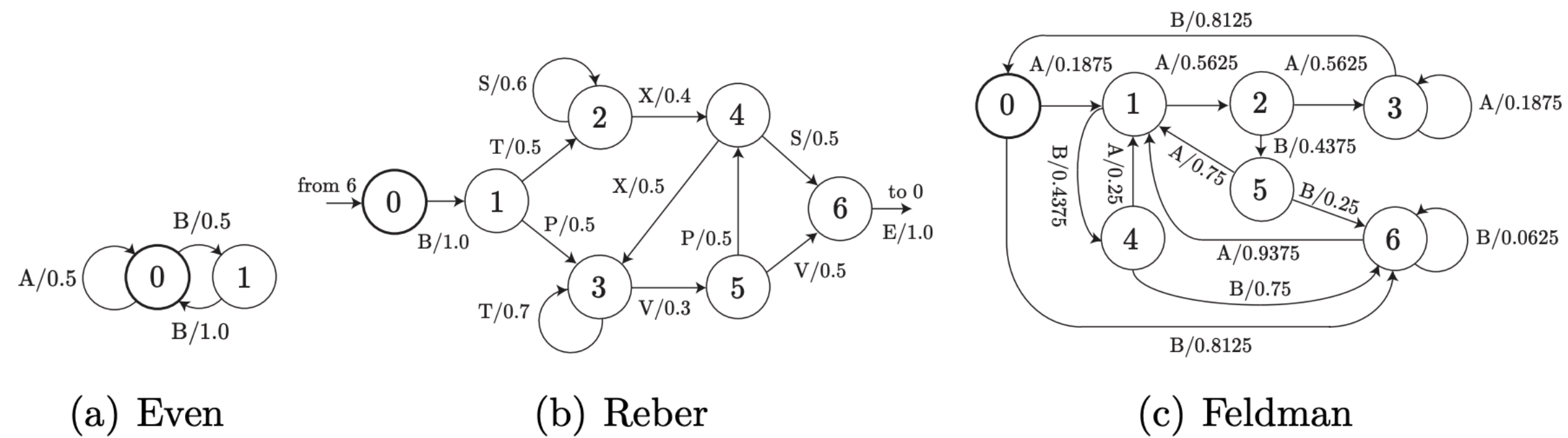
	x_0	x_1
S_1	S_3	S_2
S_2	S_2	S_3
S_3	S_1	S_4
S_4		
\vdots		



Place a *hierarchical nonparametric prior* on the state transition function

Probabilistic Deterministic Infinite Automata

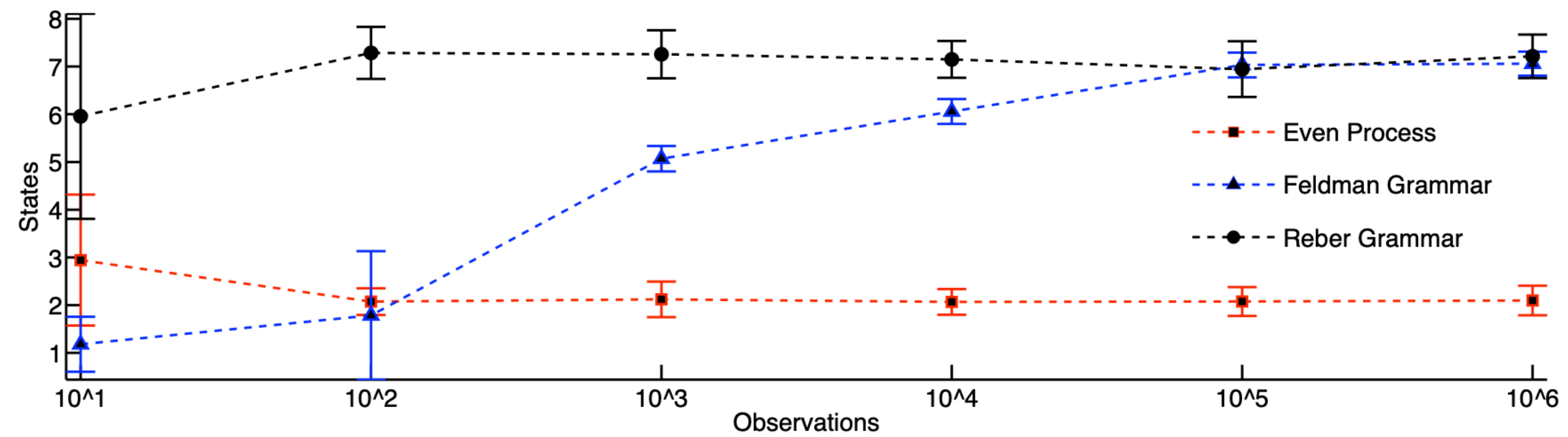
Synthetic Grammars



(a) Even

(b) Reber

(c) Feldman



(d) Posterior marginal PDIA state cardinality distribution

Probabilistic Deterministic Infinite Automata

Character-level Language Modelling

	PDIA	PDIA-MAP	HMM-EM	bigram	trigram	4-gram	5-gram	6-gram	SSM
AIW	5.13	5.46	7.89	9.71	6.45	5.13	4.80	4.69	4.78
	365.6	379	52	28	382	2,023	5,592	10,838	19,358

4-gram level accuracy with number of parameters comparable to trigram

Probabilistic Deterministic Infinite Automata

Character-level Language Modelling

	PDIA	PDIA-MAP	HMM-EM	bigram	trigram	4-gram	5-gram	6-gram	SSM
AIW	5.13	5.46	7.89	9.71	6.45	5.13	4.80	4.69	4.78
	365.6	379	52	28	382	2,023	5,592	10,838	19,358

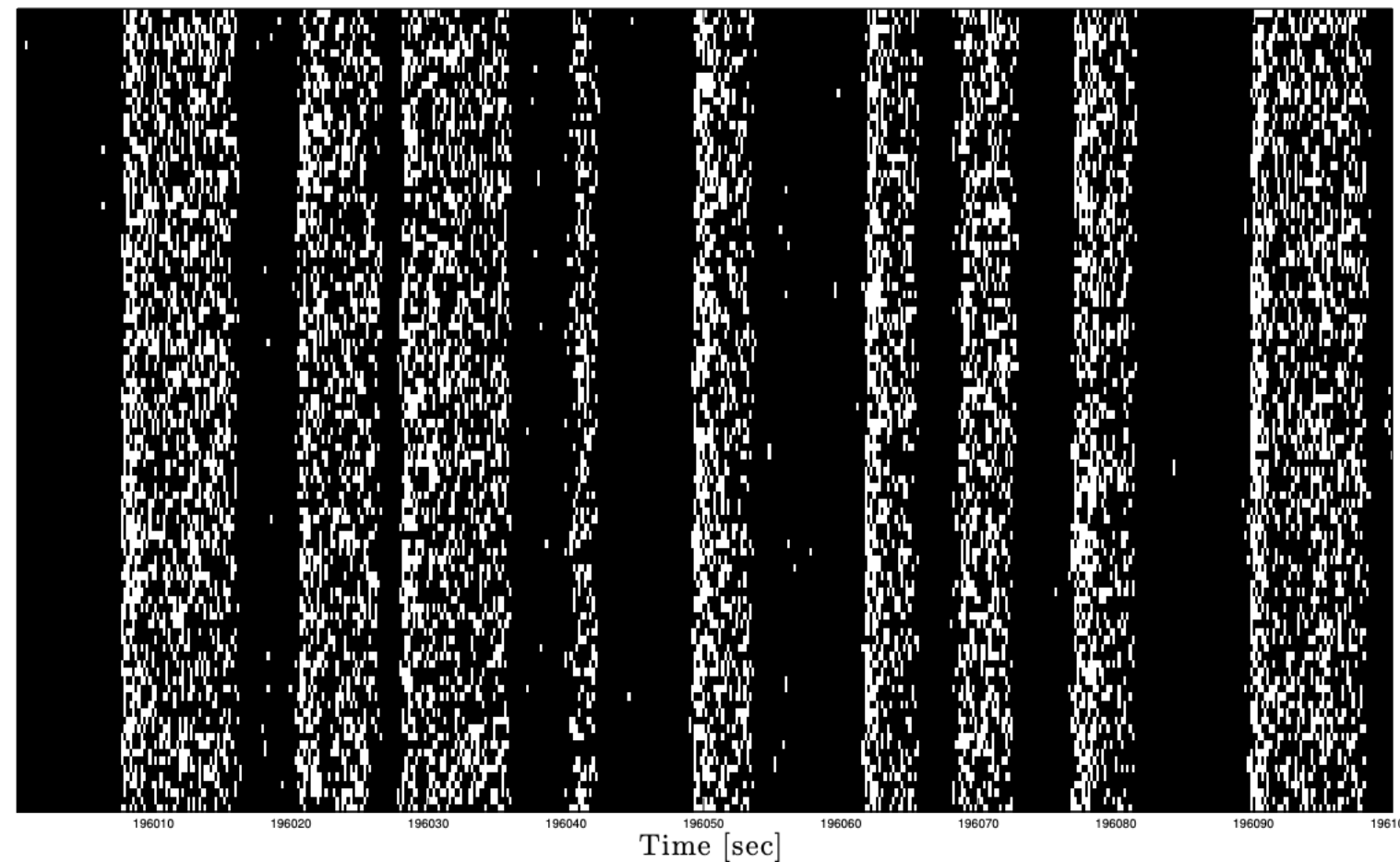
4-gram level accuracy with number of parameters comparable to trigram

what a mushroom very softly have the the way either little about a deal she what to kept i to when b...
what you her and took when pim bill alice himself ignvy conversationer after treat eye going very to...
seside must upon to the a othering in for the the i of i him of hrisall a is either mock turtle and...
however nor rats come perself for everywheelsome something ll in hoor of but her said you heople was...
whistled inqueer hersonate it doing daiek the ll the she be away the the queen than of that miss pea...

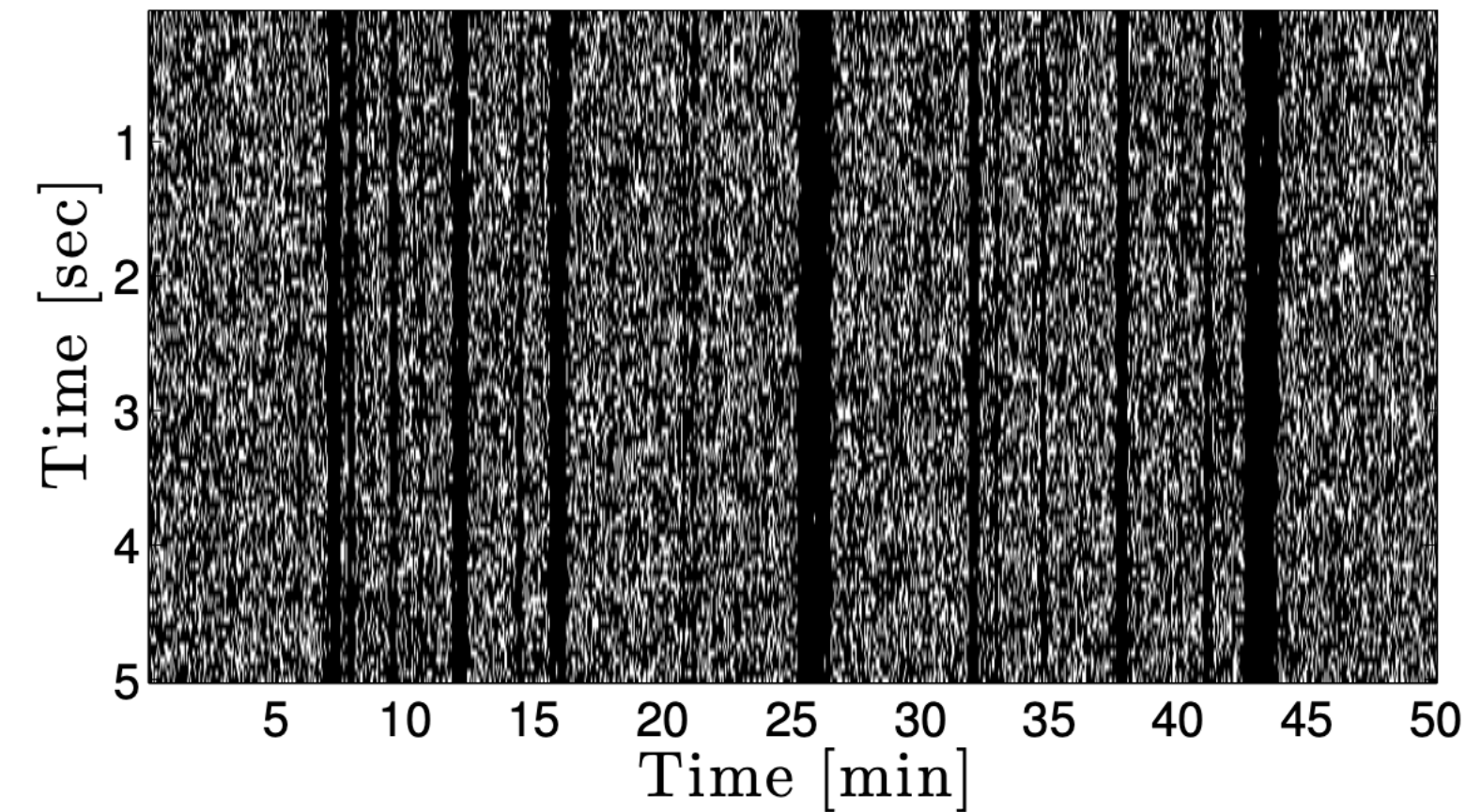
Probabilistic Deterministic Infinite Automata

Single Neuron Recordings

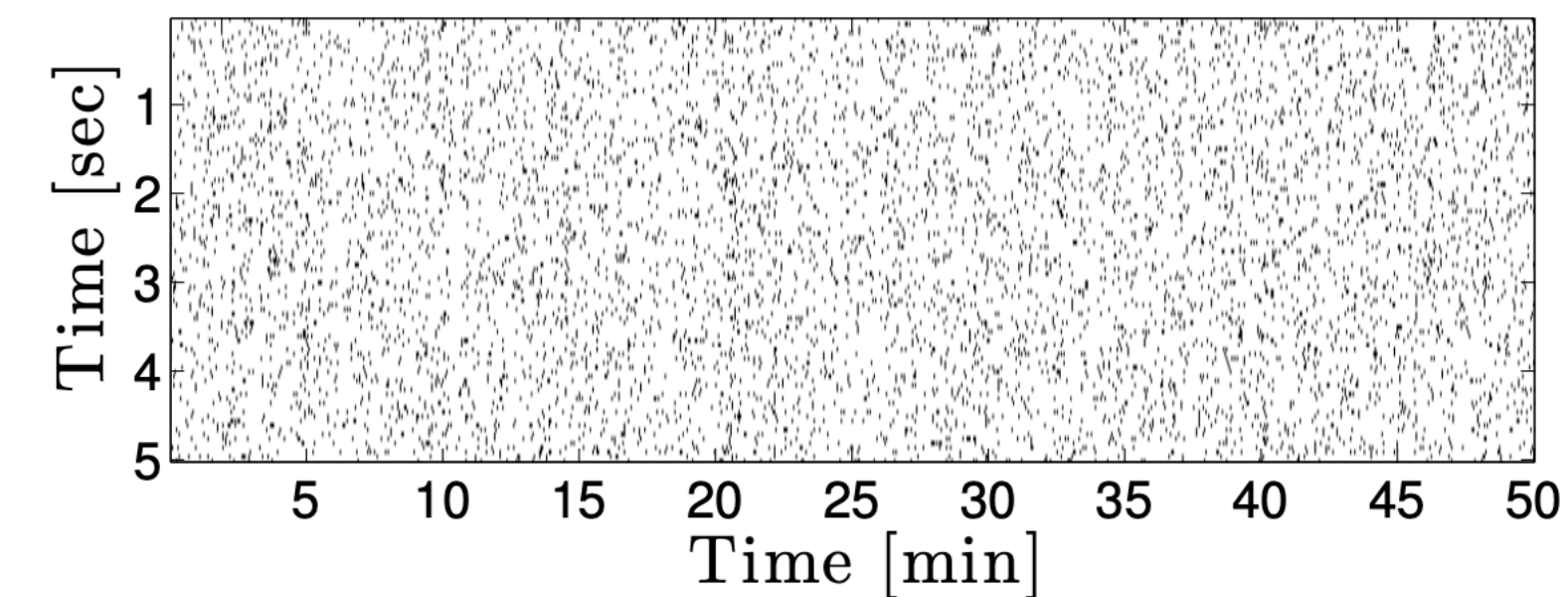
Experiment



PDIA

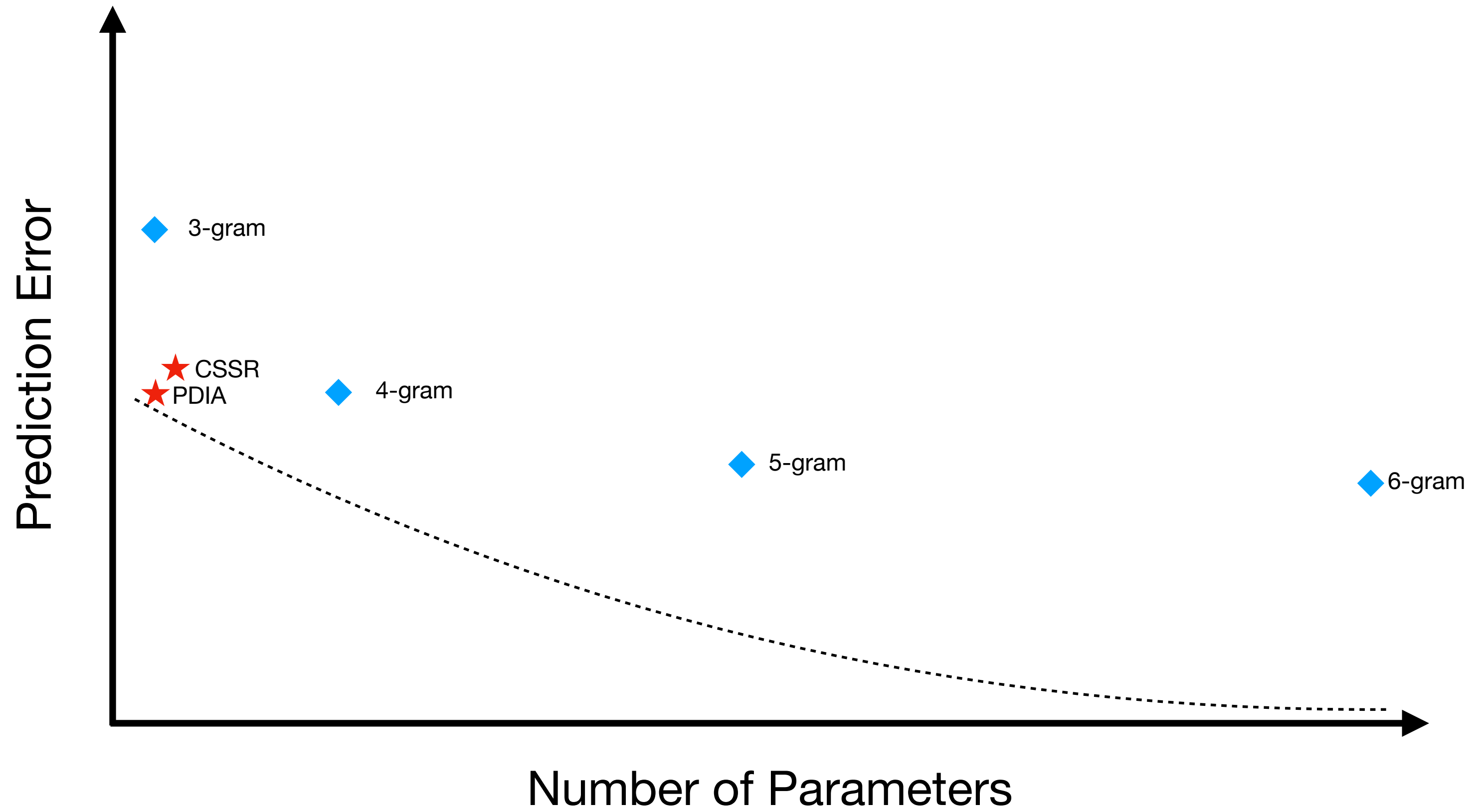


GLM(100)



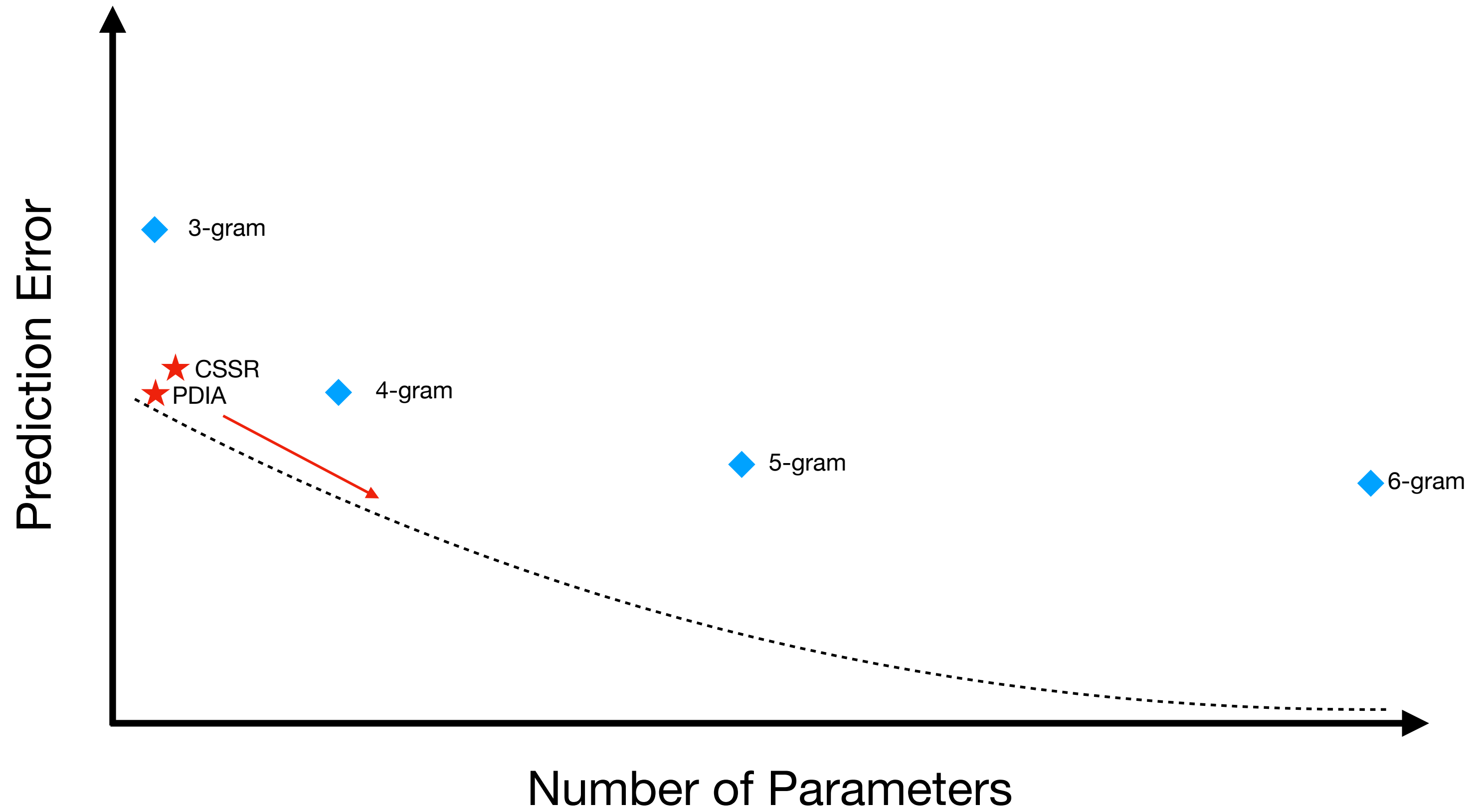
Probabilistic Deterministic Infinite Automata

Scaling Up?



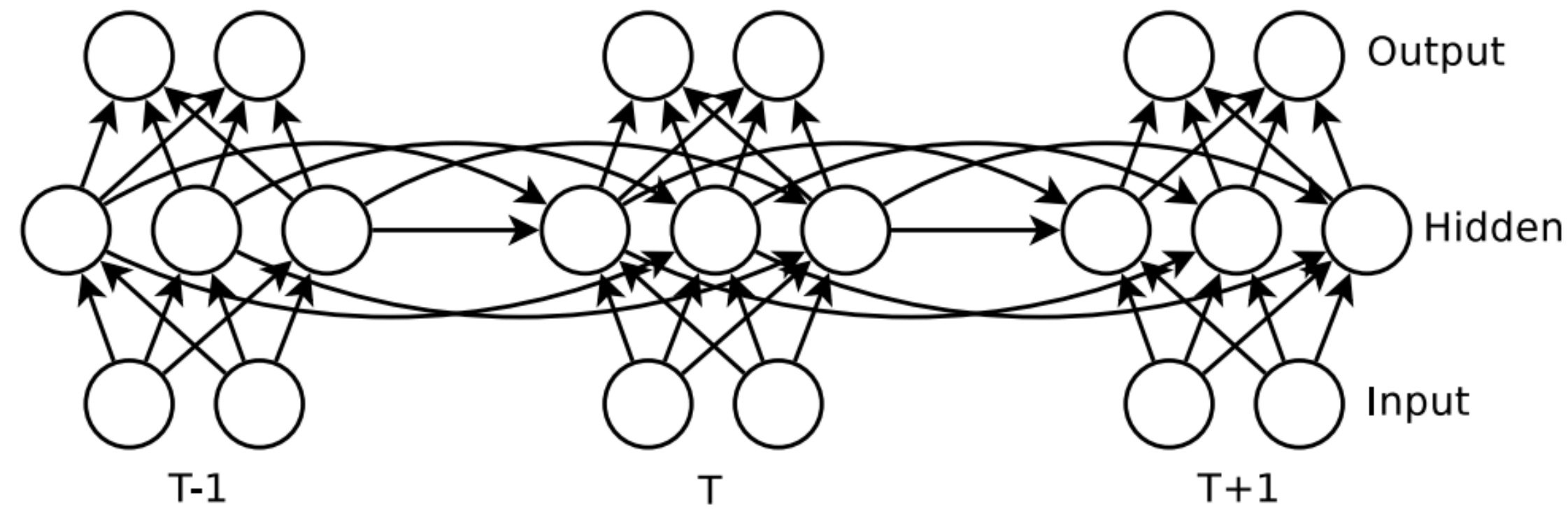
Probabilistic Deterministic Infinite Automata

Scaling Up?



Neural Language Models

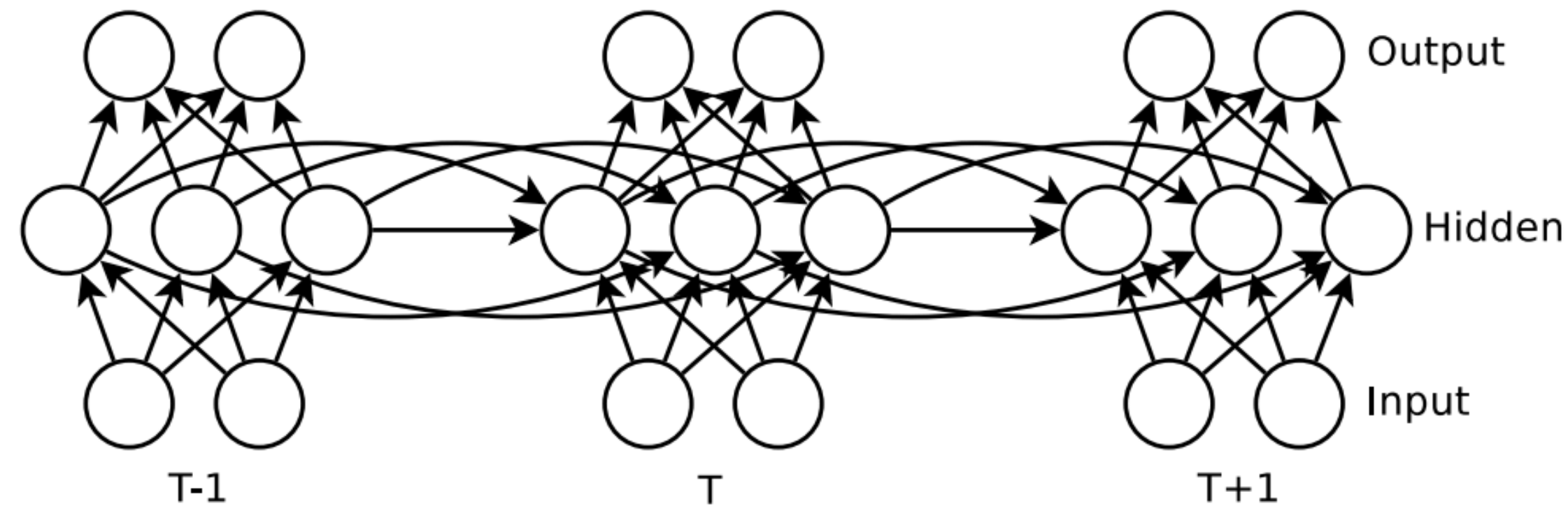
Recurrent Neural Networks



DATA SET	MEMOIZER	PAQ	MRNN	MRNN (FULL SET)
WIKI	1.66	1.51	1.60 (1.53)	1.55 (1.54)
NYT	1.49	1.38	1.48 (1.44)	1.47 (1.46)
ML	1.33	1.22	1.31 (1.27)	

Neural Language Models

Recurrent Neural Networks



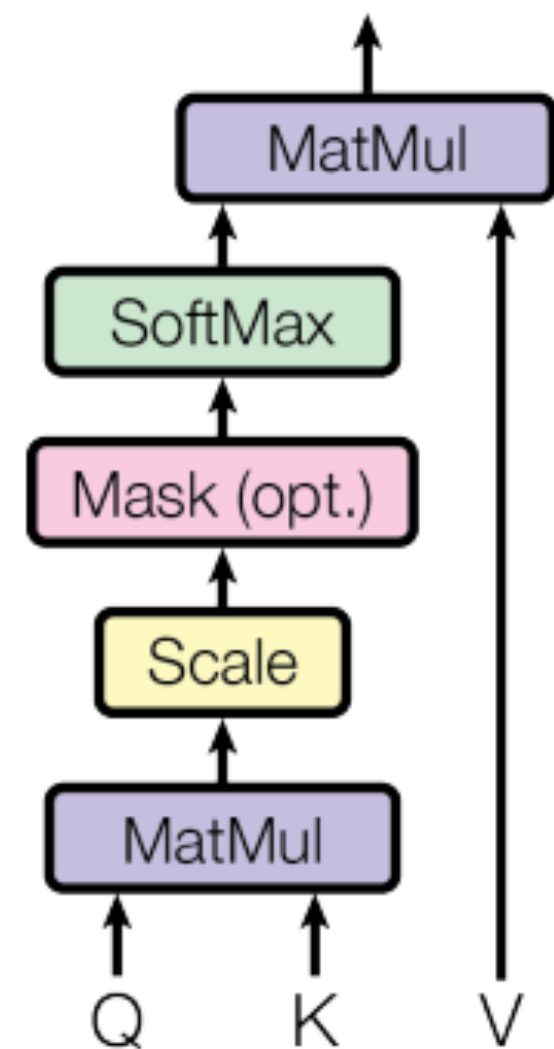
DATA SET	MEMOIZER	PAQ	MRNN	MRNN (FULL SET)
WIKI	1.66	1.51	1.60 (1.53)	1.55 (1.54)
NYT	1.49	1.38	1.48 (1.44)	1.47 (1.46)
ML	1.33	1.22	1.31 (1.27)	

The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pastured with consistent street forests were incorporated by the 15th century BE. In 1996 the primary rapford undergoes an effort that the reserve conditioning, written into Jewish cities, sleepers to incorporate the .St Eurasia that activates the population. Mar??a Nationale, Kelli, Zedlat-Dukastoe, Florendon, Ptu's thought is. To adapt in most parts of North America, the dynamic fairy Dan please believes, the free speech are much related to the

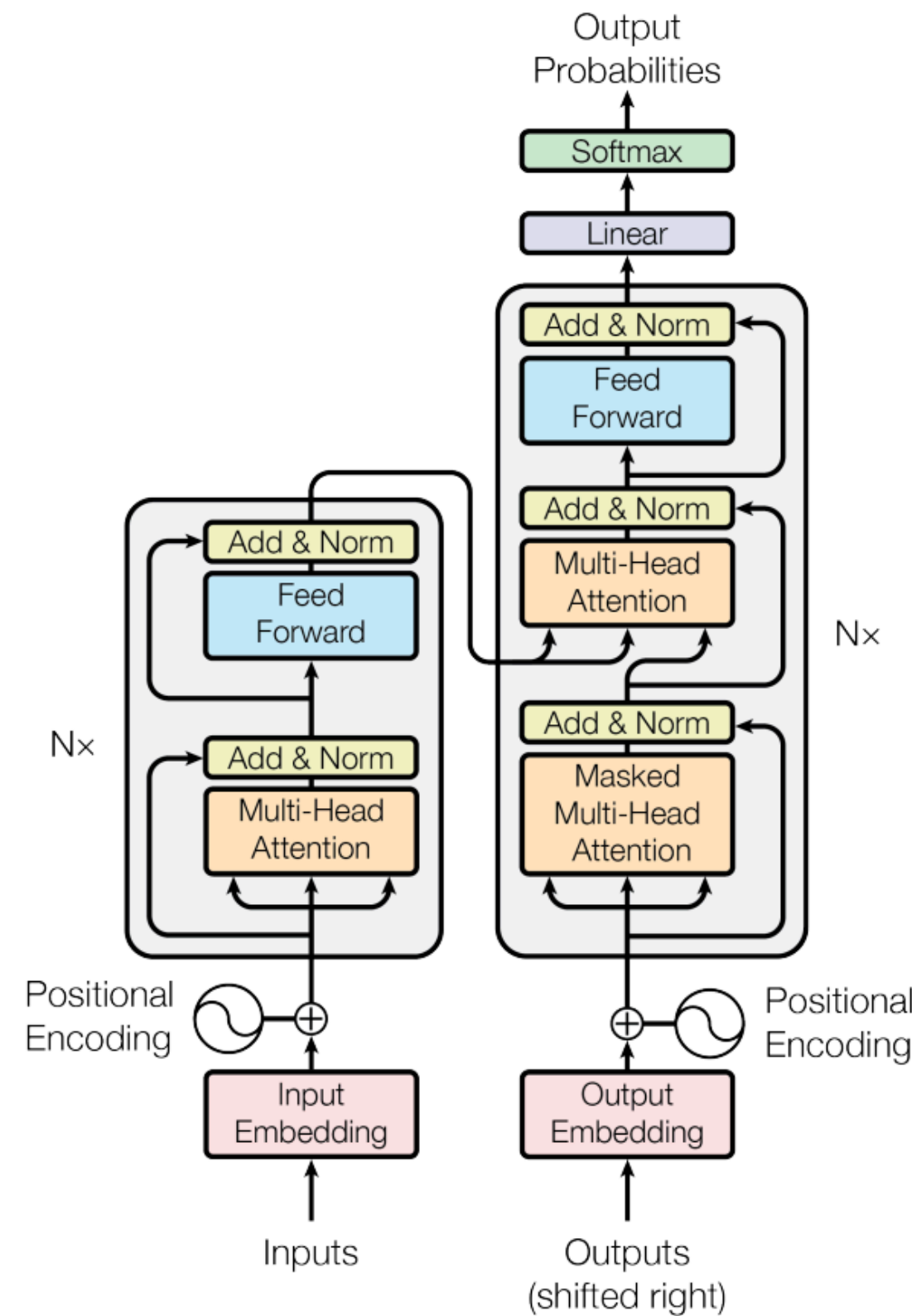
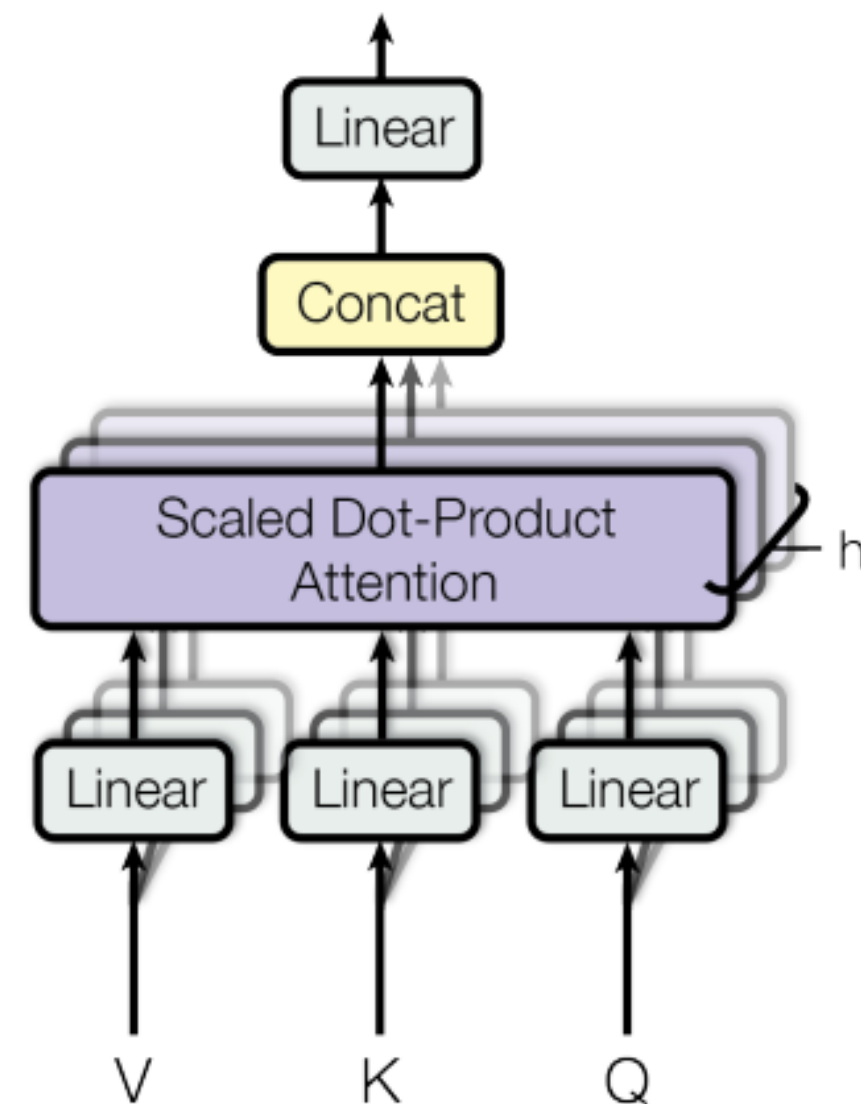
Neural Language Models

Generative Pretrained Transformers

Scaled Dot-Product Attention



Multi-Head Attention



A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, NeurIPS (2017)

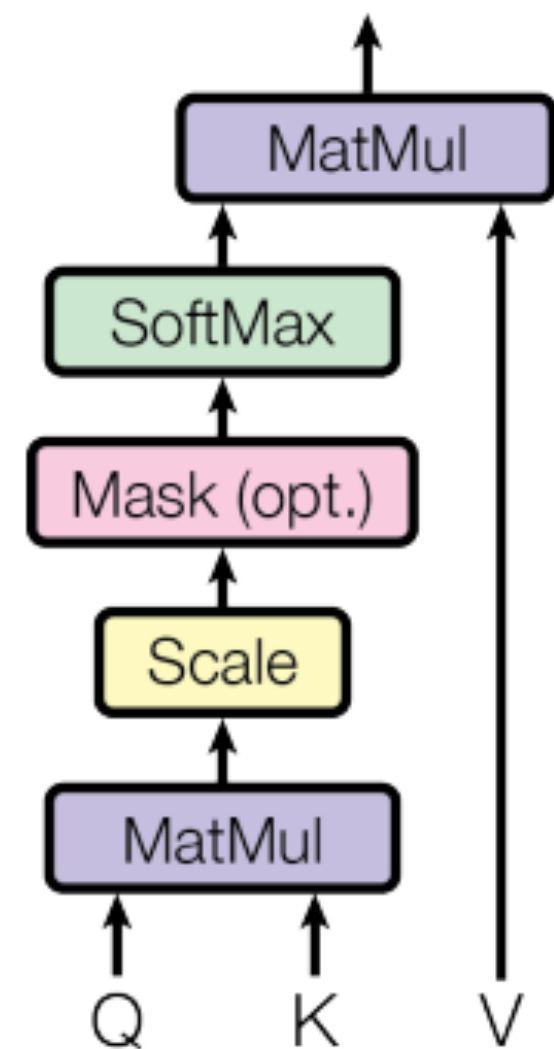
A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, NeurIPS (2018)

S. Bubeck, V. Chandrasekaran *et al.* (2023)

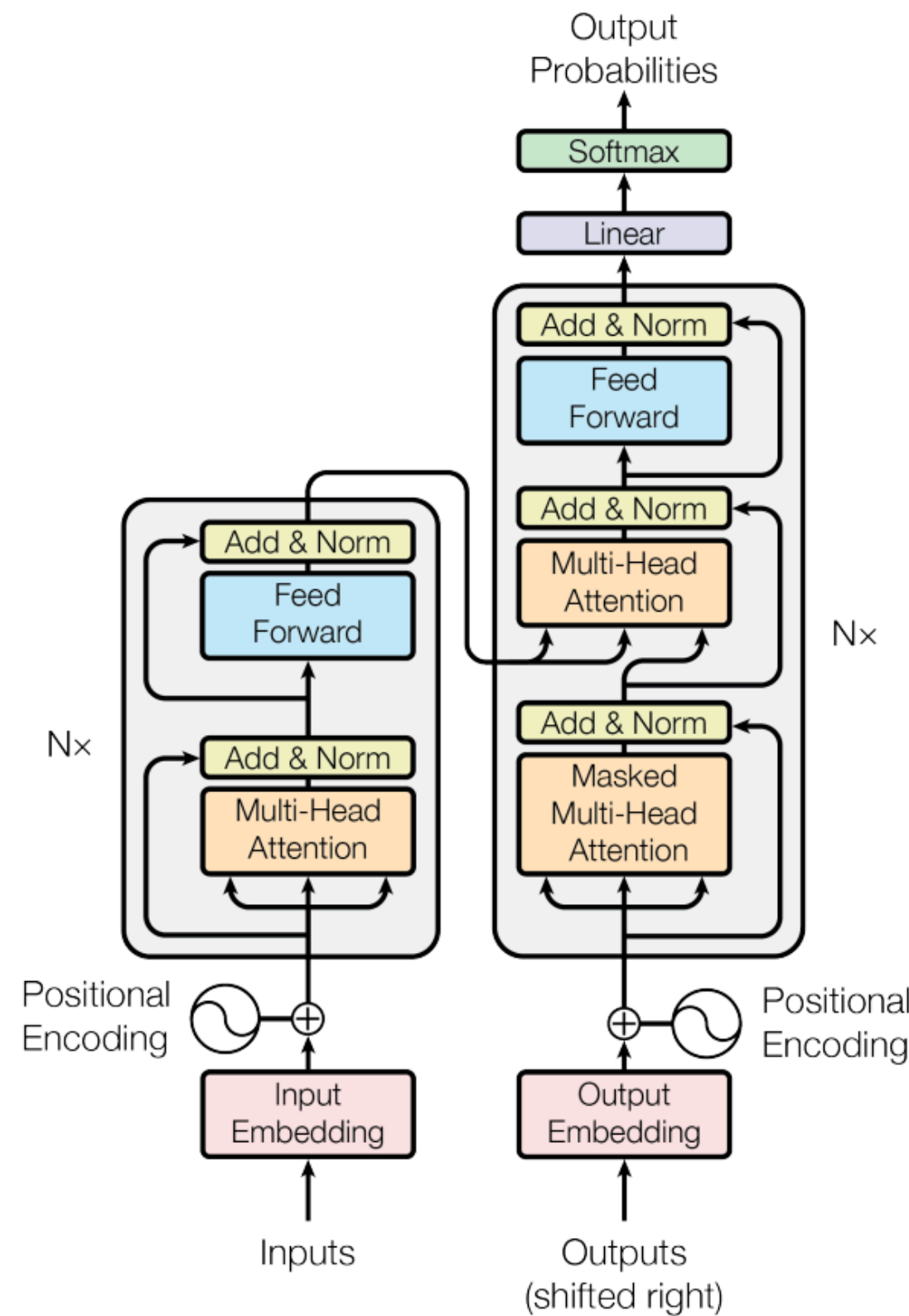
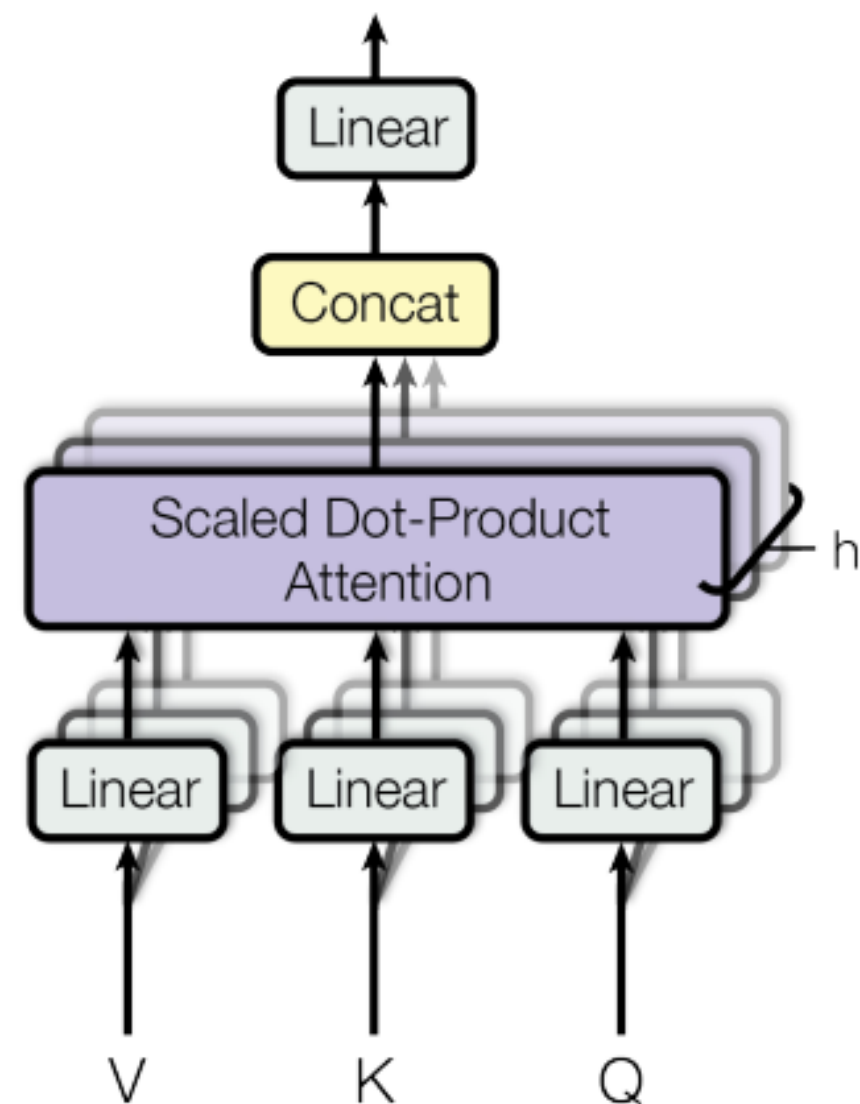
Neural Language Models

Generative Pretrained Transformers

Scaled Dot-Product Attention



Multi-Head Attention



Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

Yes, I think I can,
 Though it might take a clever plan.
 I'll start by noting Euclid's proof,
 Which shows that primes aren't just aloof.
 Assume we have a finite list,
 Of primes, and that none have been missed.
 Multiply them all together,
 And add one, just to be clever.
 The result will be a number,
 That has no prime factors, oh wonder!
 But every number has a prime,
 That divides it, it's just a matter of time.
 So we've found a contradiction,
 And our finite list needs eviction.
 There must be infinitely many primes,
 And that's the end of my rhyming lines.s

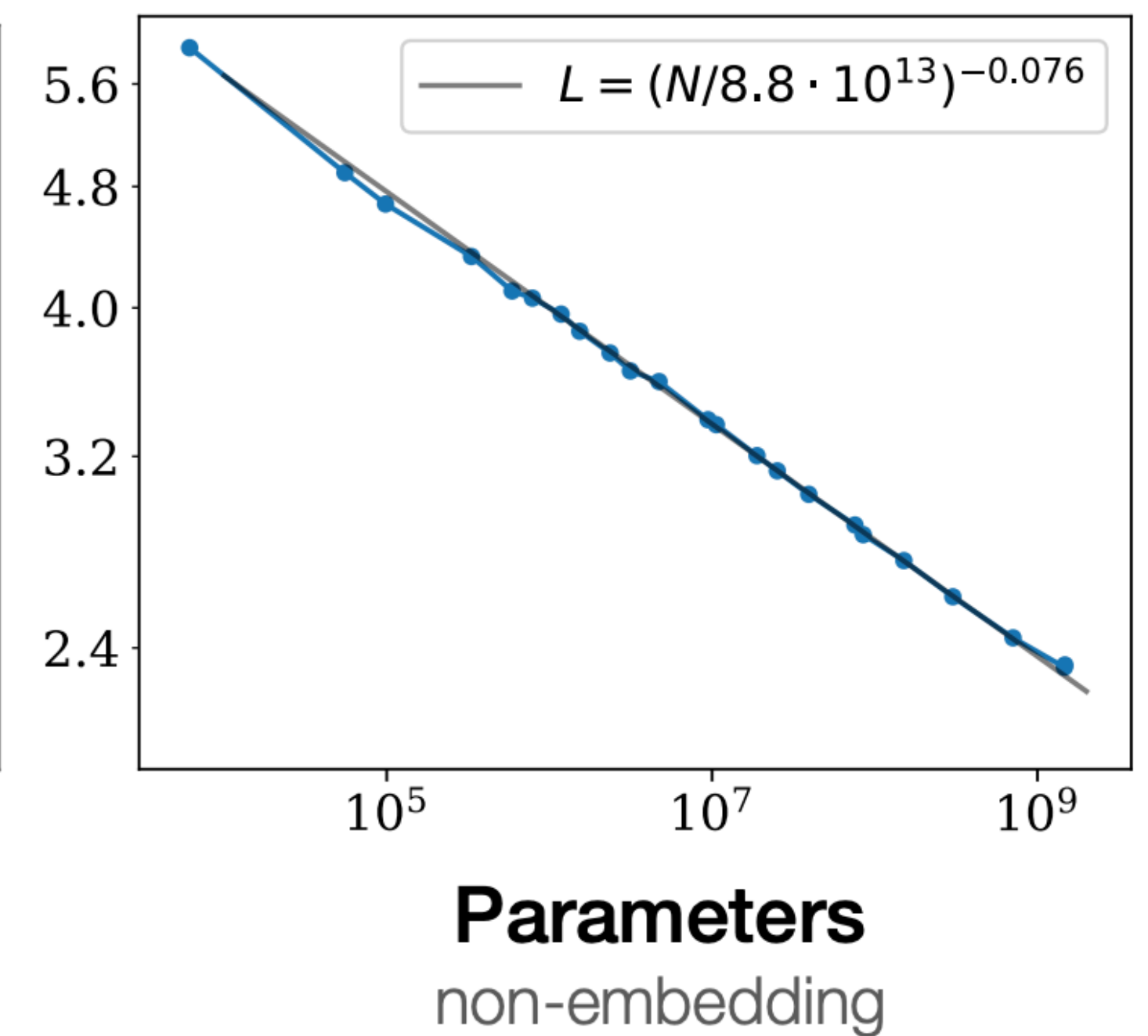
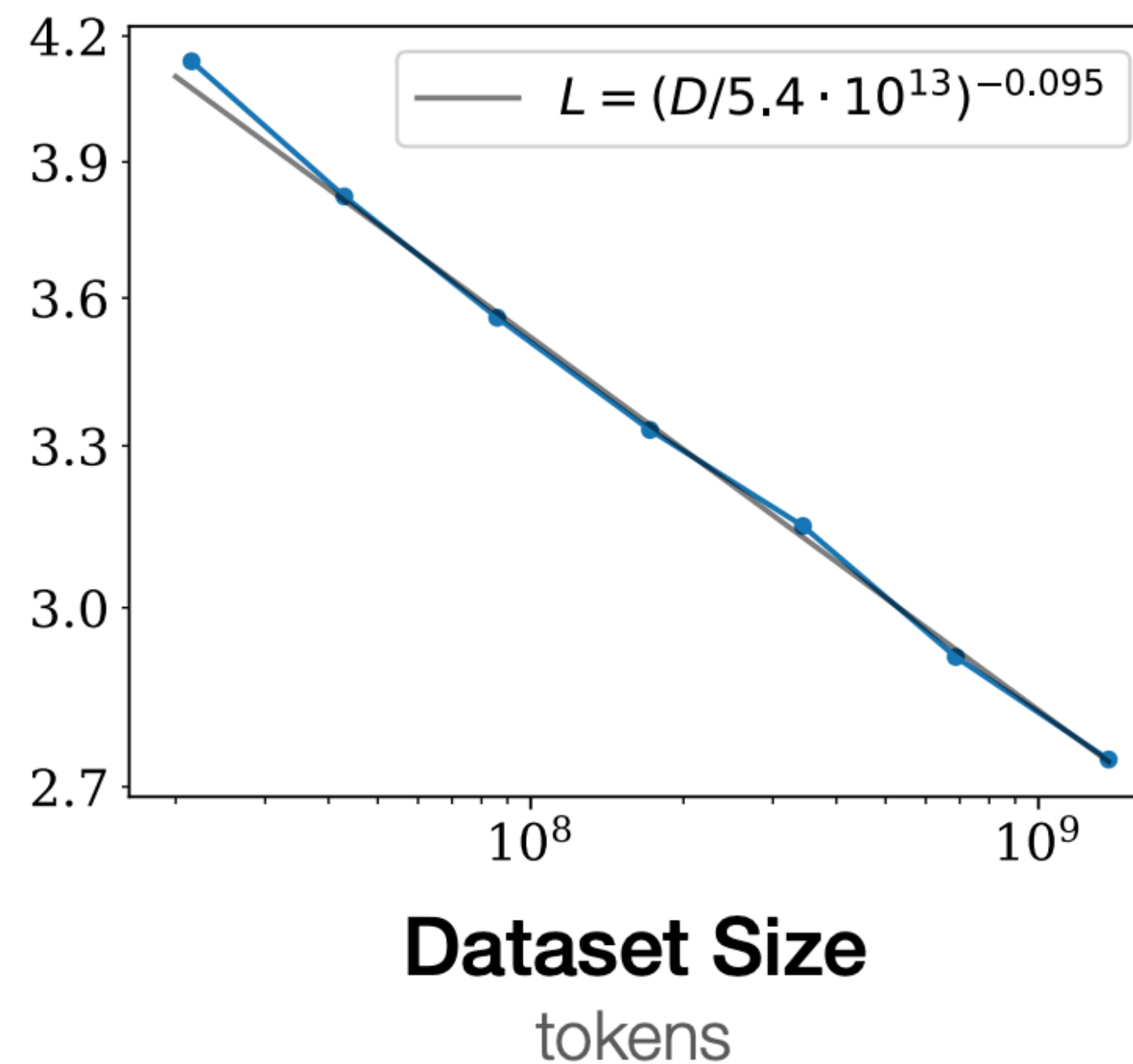
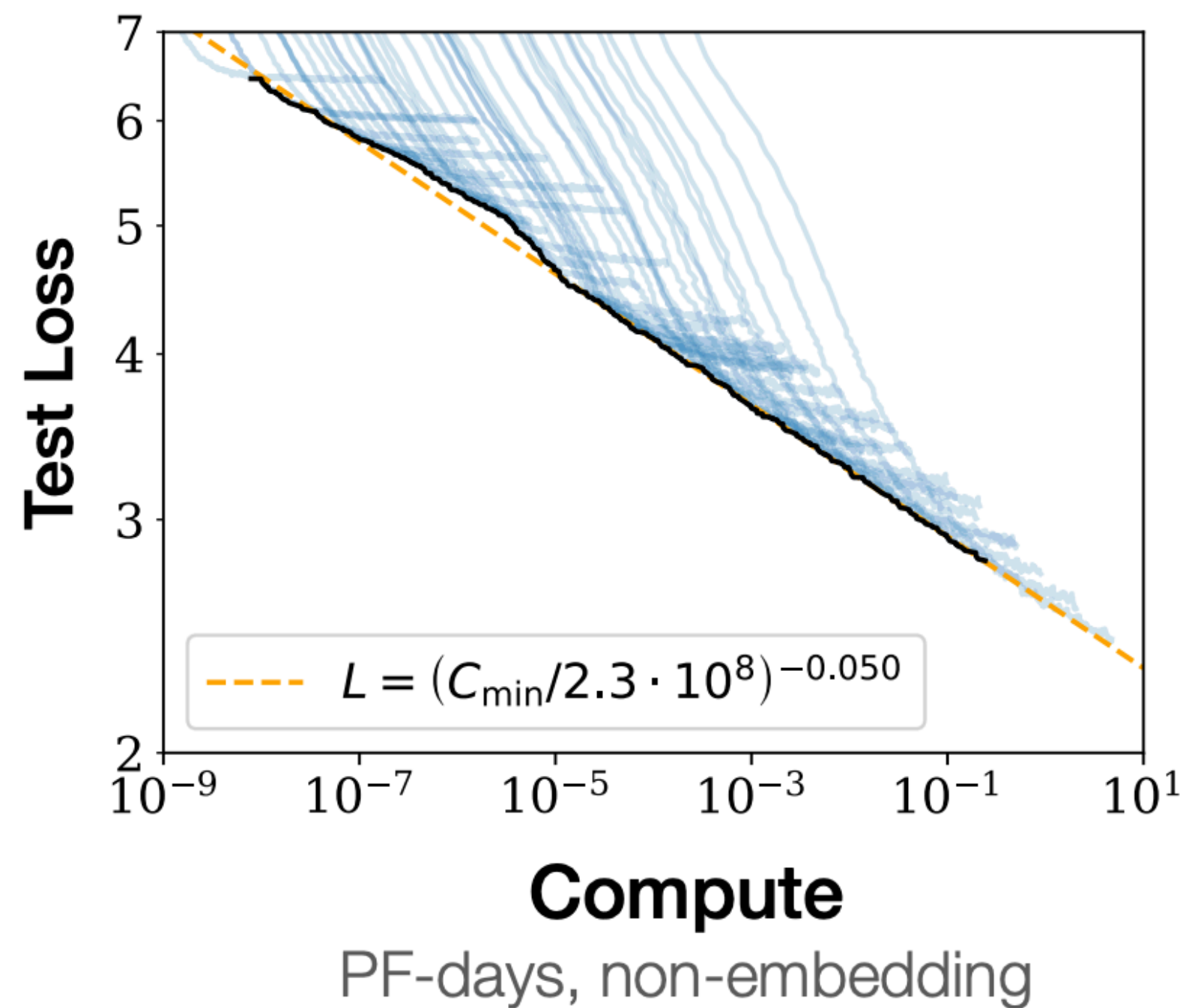
A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, NeurIPS (2017)

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, NeurIPS (2018)

S. Bubeck, V. Chandrasekaran *et al.* (2023)

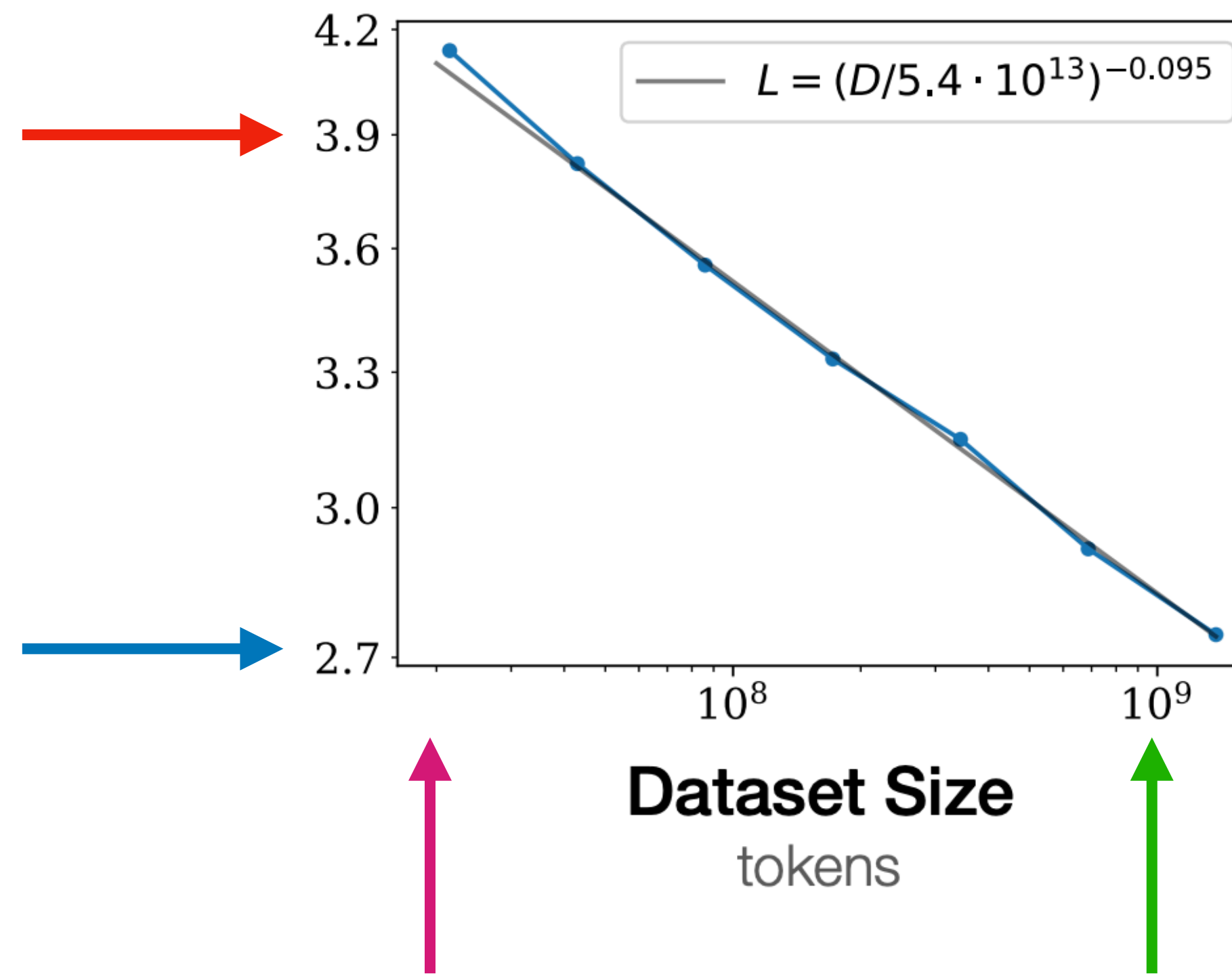
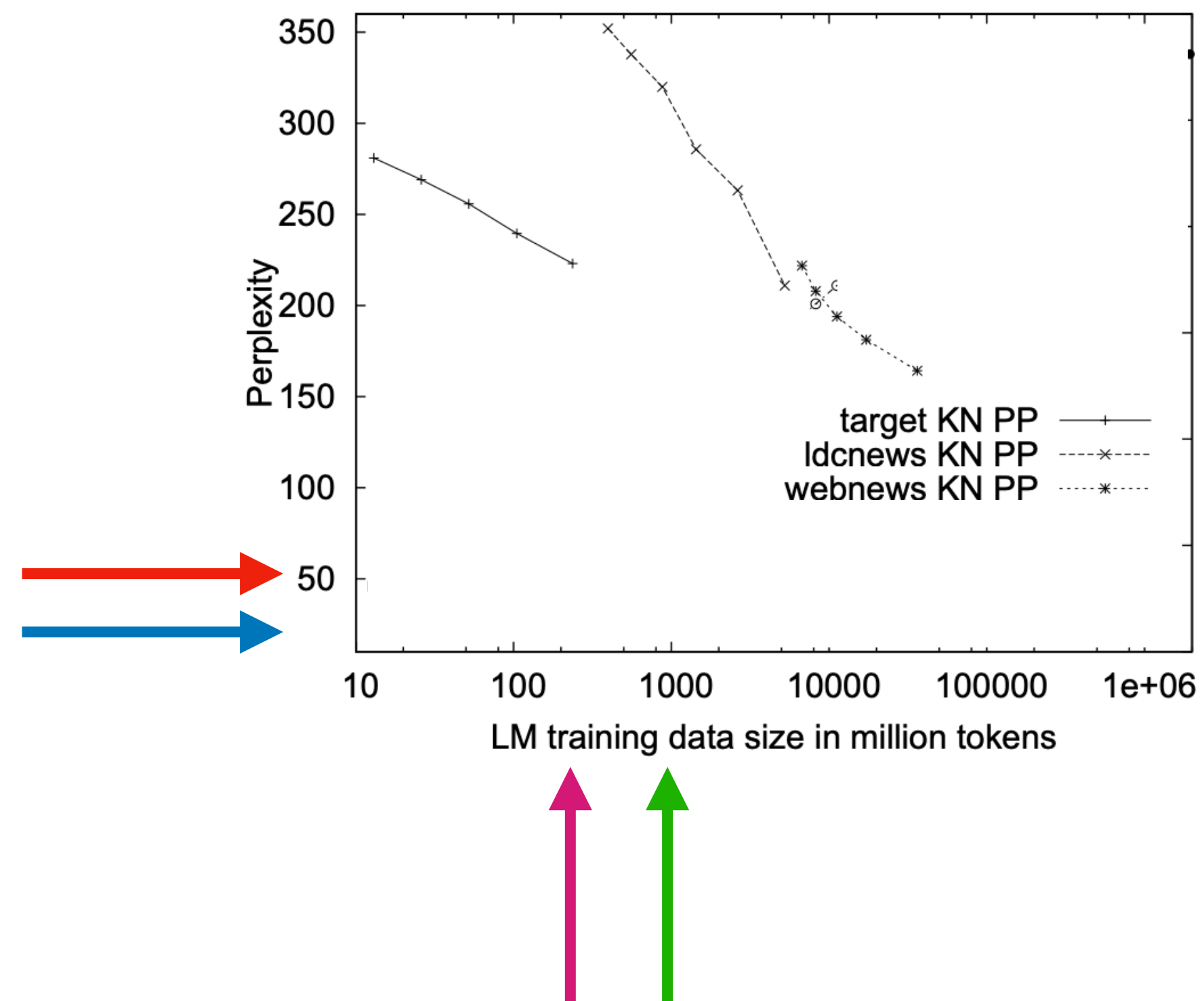
Neural Language Models

Power Law Scaling Laws



Neural Language Models

Power Law Scaling Laws

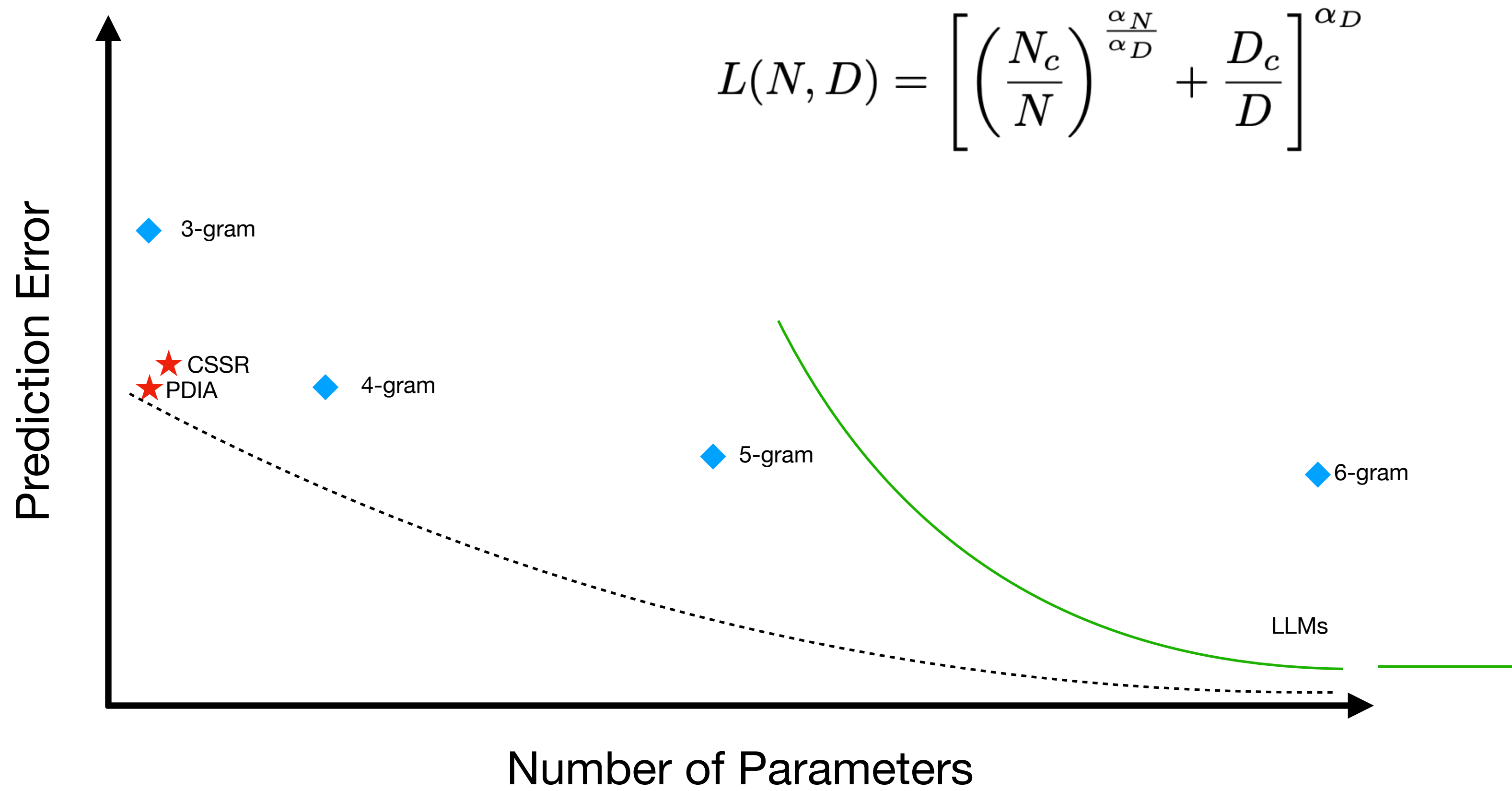


J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei (2020)

T. Brants, A. C. Popat, P. Xu, F. J. Ochs, J. Dean, EMNLP (2007)

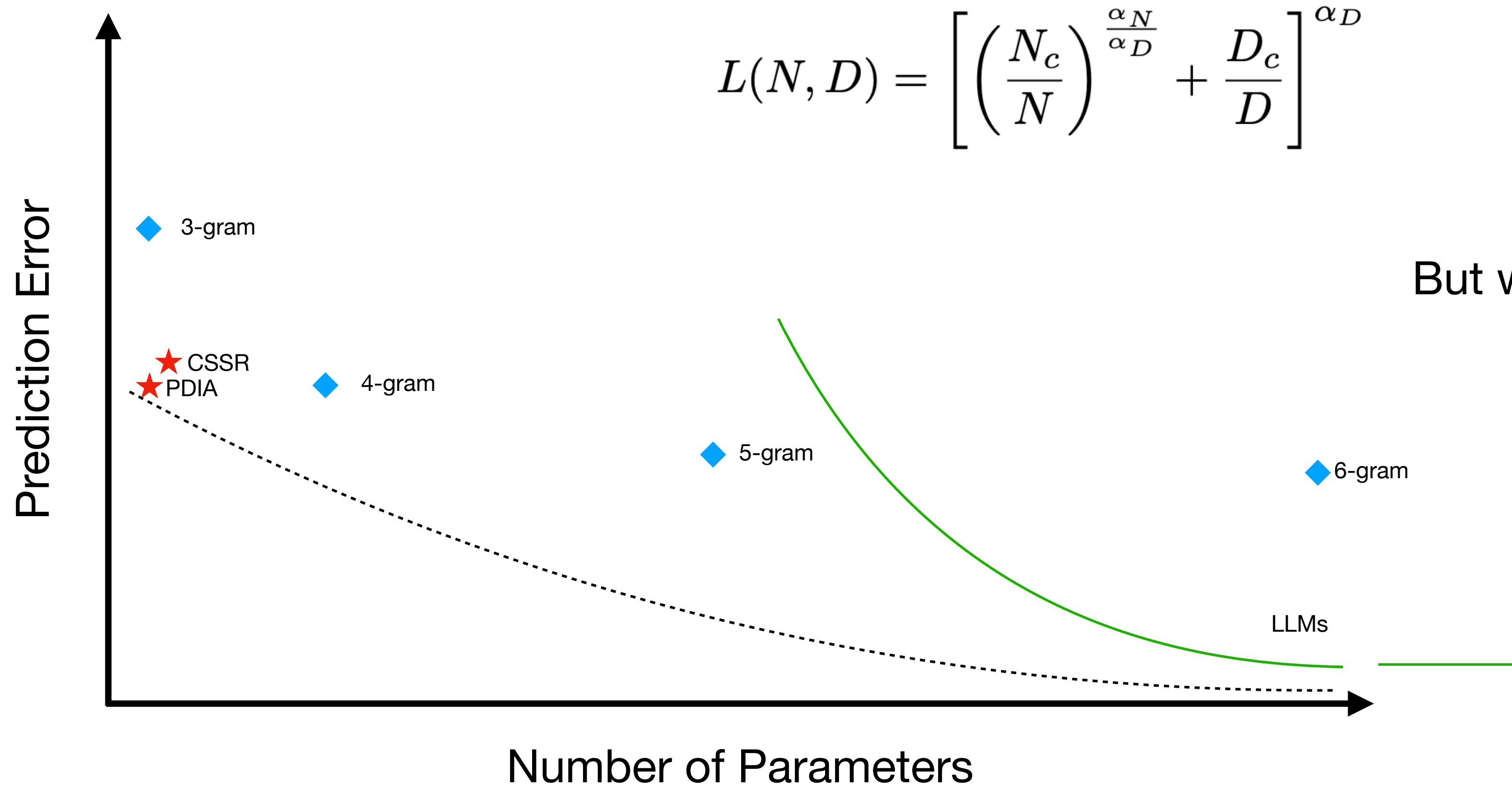
Neural Language Models

Scaling is all you need?

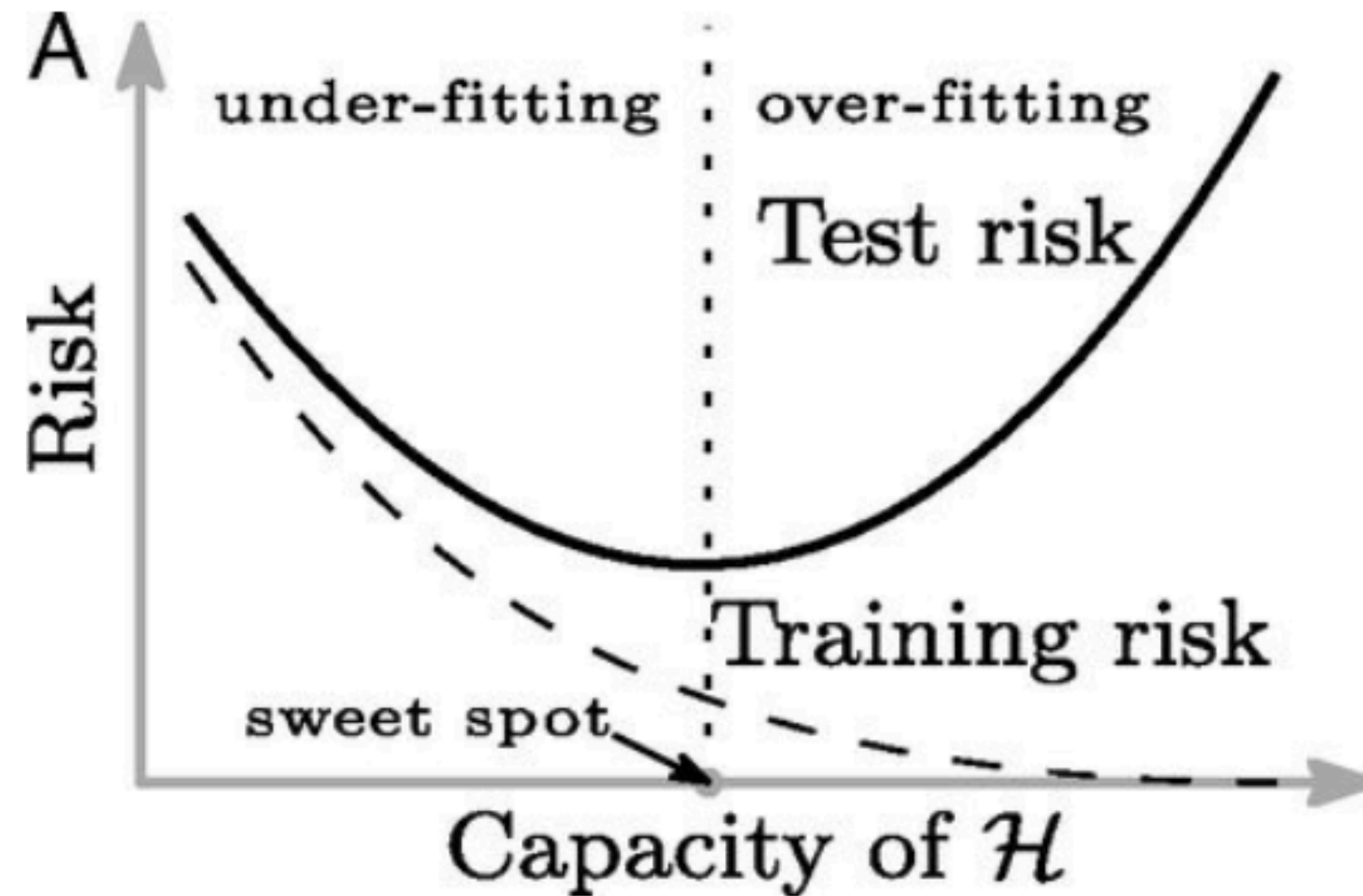


Neural Language Models

Scaling is all you need?

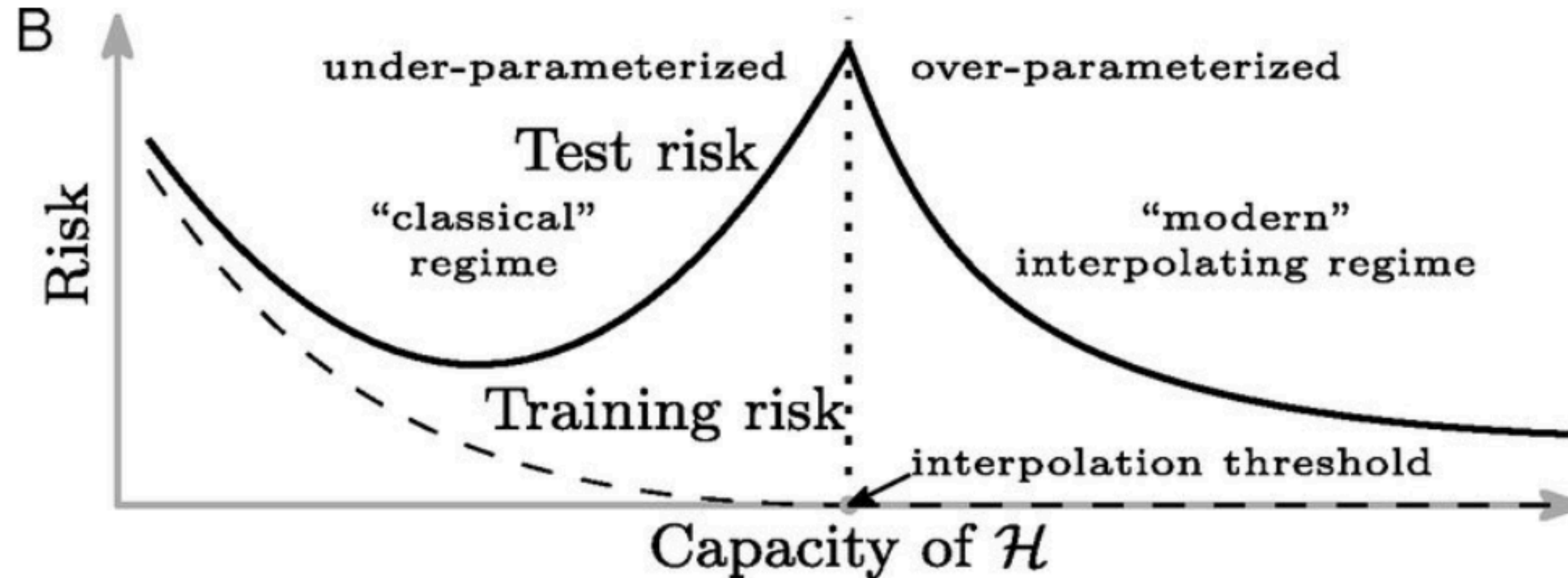


Double Descent



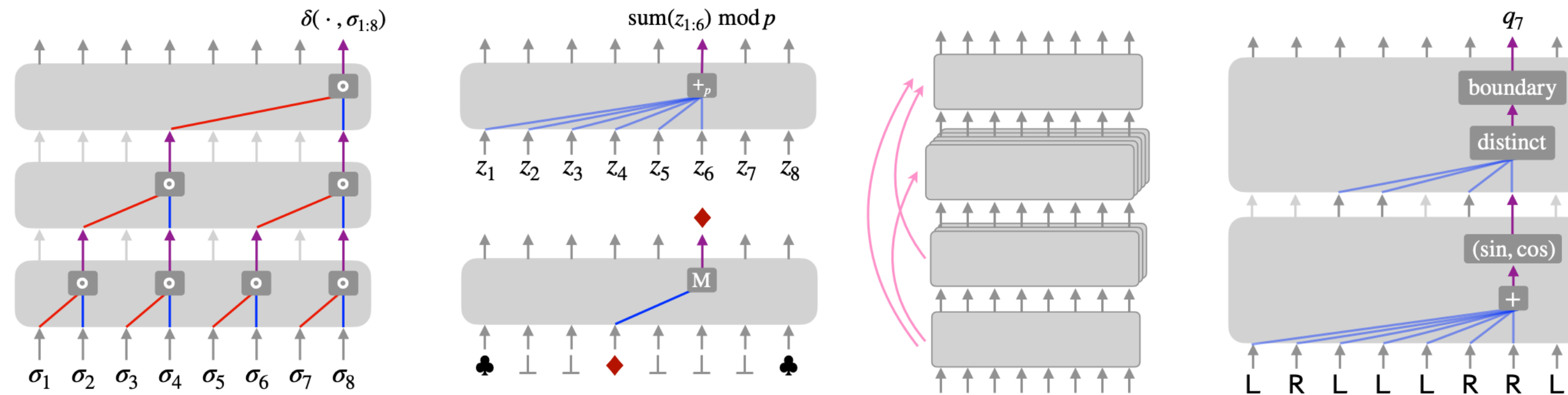
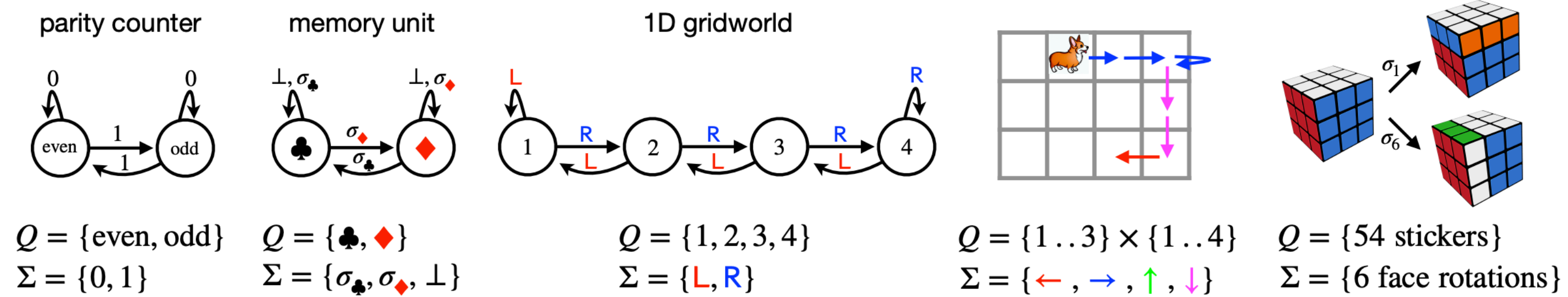
- Bayesian/MDL/Bias-variance story holds up until the model has the capacity to memorize the data

Double Descent



- Bayesian/MDL/Bias-variance story holds up until the model has the capacity to memorize the data
- Beyond that - the extra parameters *help* prevent overfitting!

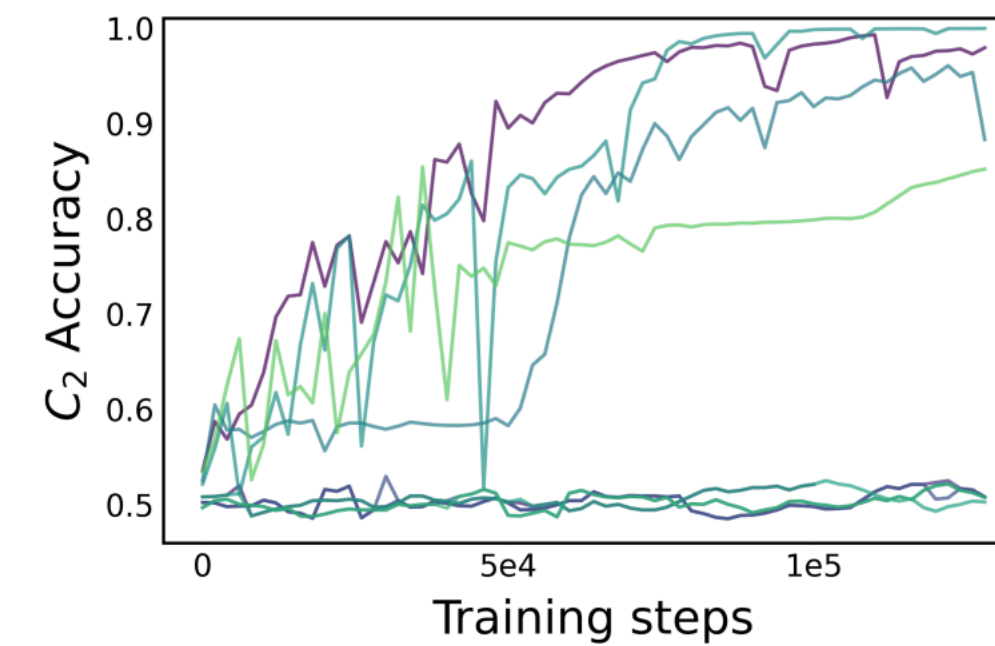
Can Transformers Learn Simple Automata?



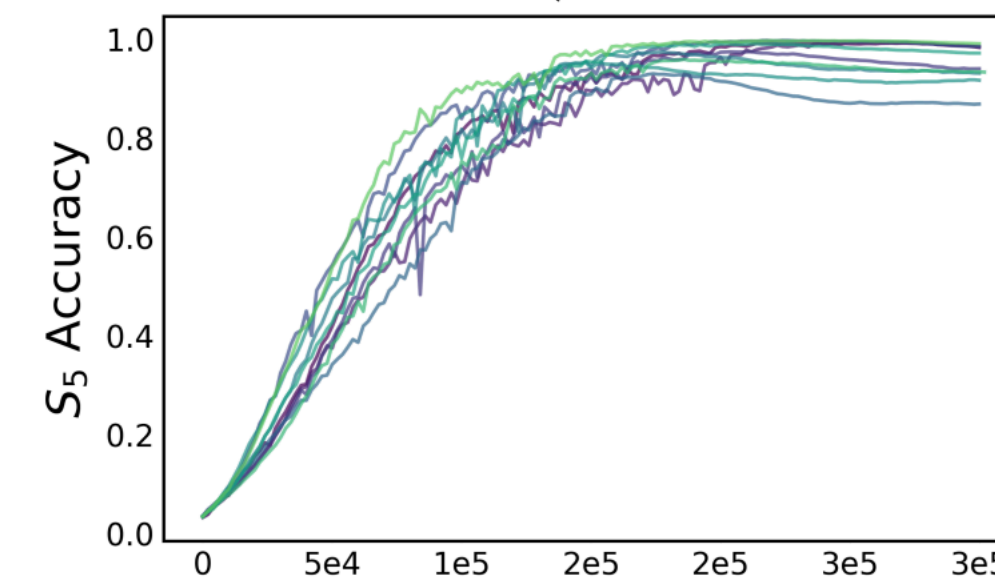
Semiautomata on length T sequences can be implemented in depth $\log(T)$ Transformers

Can Transformers Learn Simple Automata?

	1	2	3	4	5	6	7	8	12	16
Dyck	99.3	100	100	100	100	100	100	100	100	100
Grid ₉	92.2	100	100	100	100	100	100	100	100	100
C ₂	77.6	99.8	99.9	100	100	99.5	100	99.7	100	100
C ₃	54.6	94.6	96.7	99.4	100	100	99.8	100	100	100
C ₂ ³	65.0	77.9	99.9	97.9	100	99.8	98.2	99.9	95.9	80.6
D ₆	25.4	27.2	47.4	75.2	100	100	100	100	100	100
D ₈	45.6	98.0	100	100	100	100	100	100	100	100
Q ₈	31.6	49.2	59.6	60.4	73.5	99.3	100	100	100	100
A ₅	12.5	23.1	32.5	46.7	71.2	98.8	100	100	100	100
S ₅	7.9	11.8	14.6	19.7	26.0	28.4	32.8	51.8	97.2	99.9

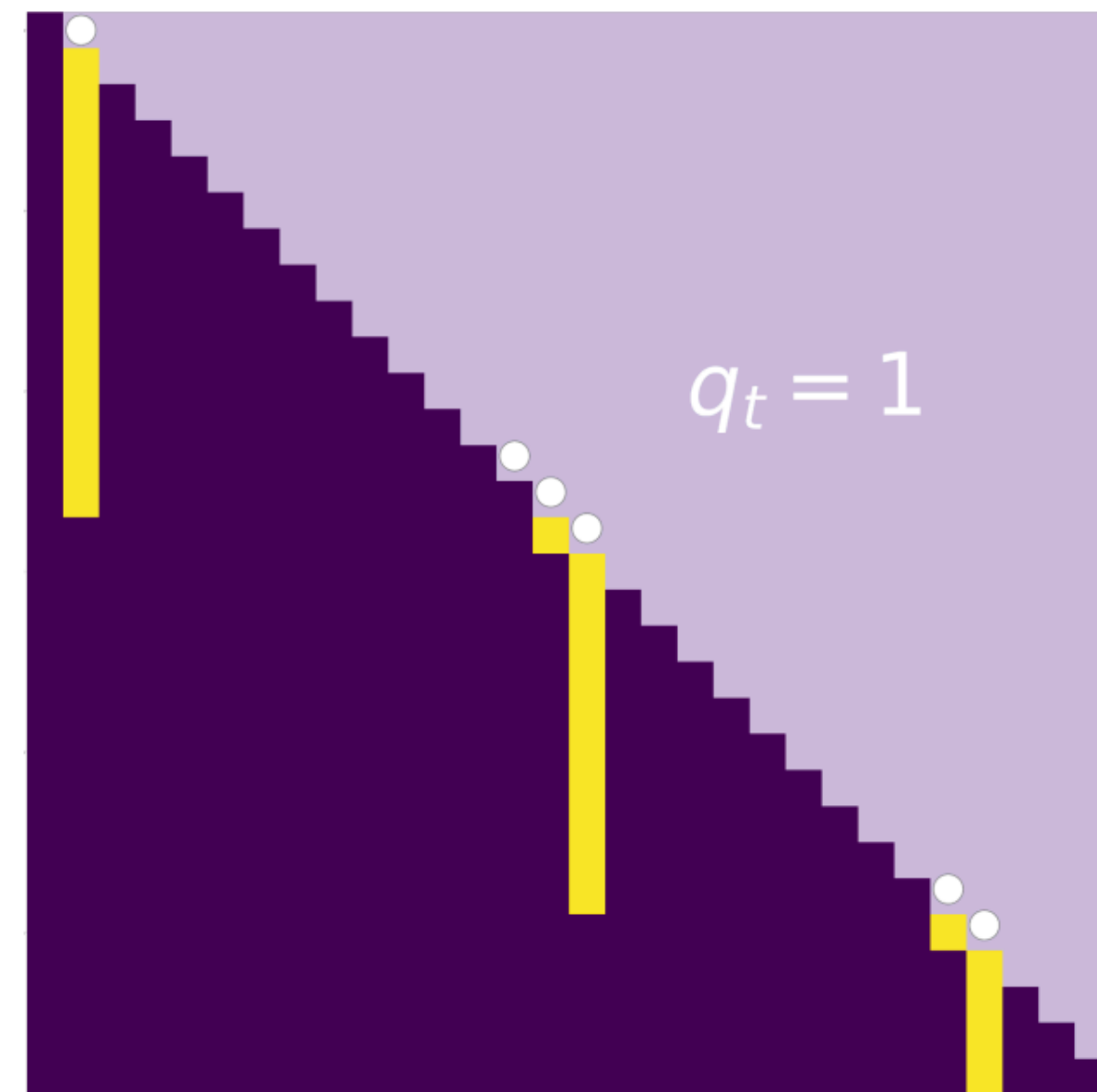


(b) Training curves for C₂ (i.e. parity; 10 replicates).

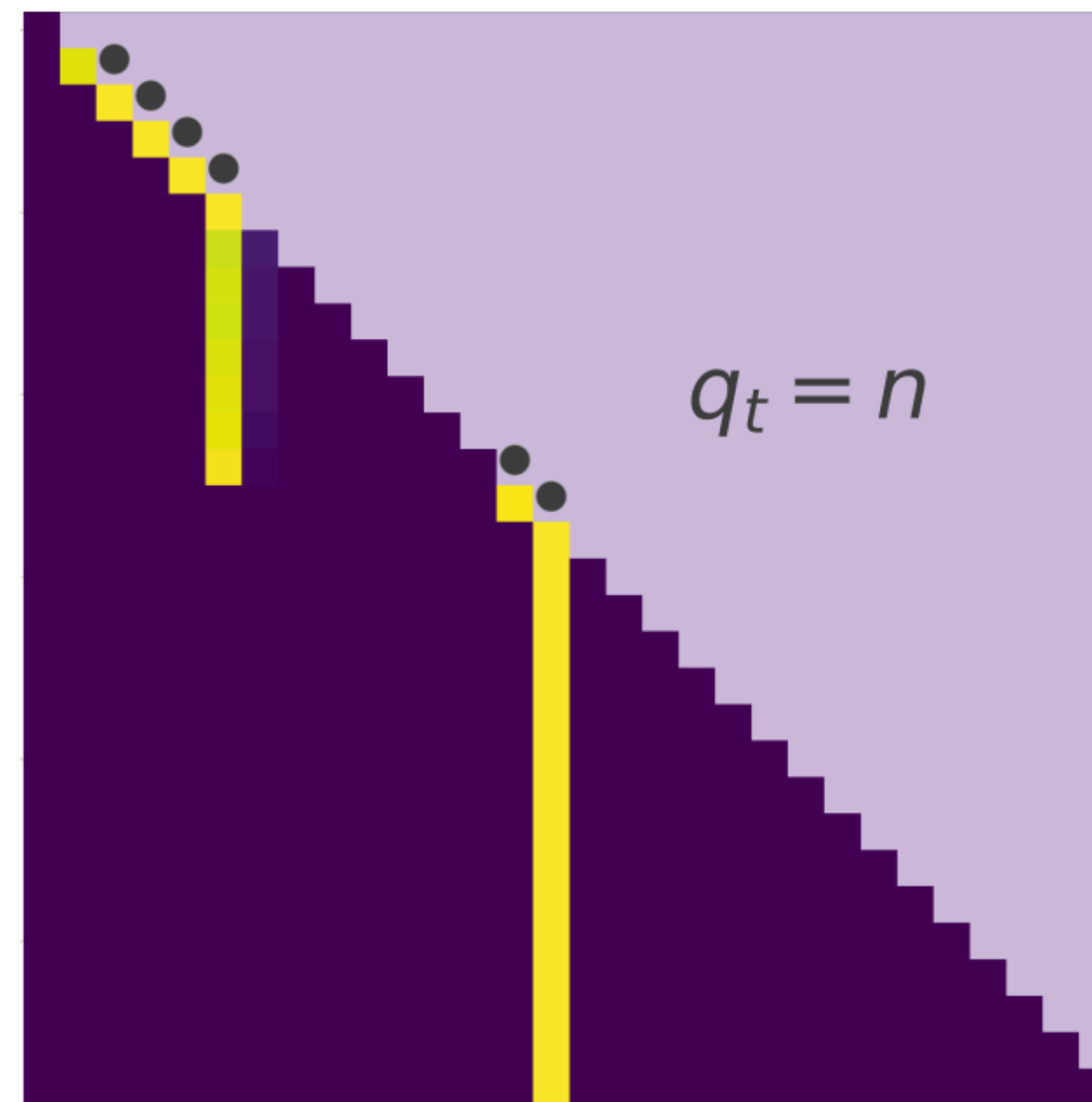


Can learn semiautomata in practice with enough depth - but with insane # of parameters

Can Transformers Learn Simple Automata?



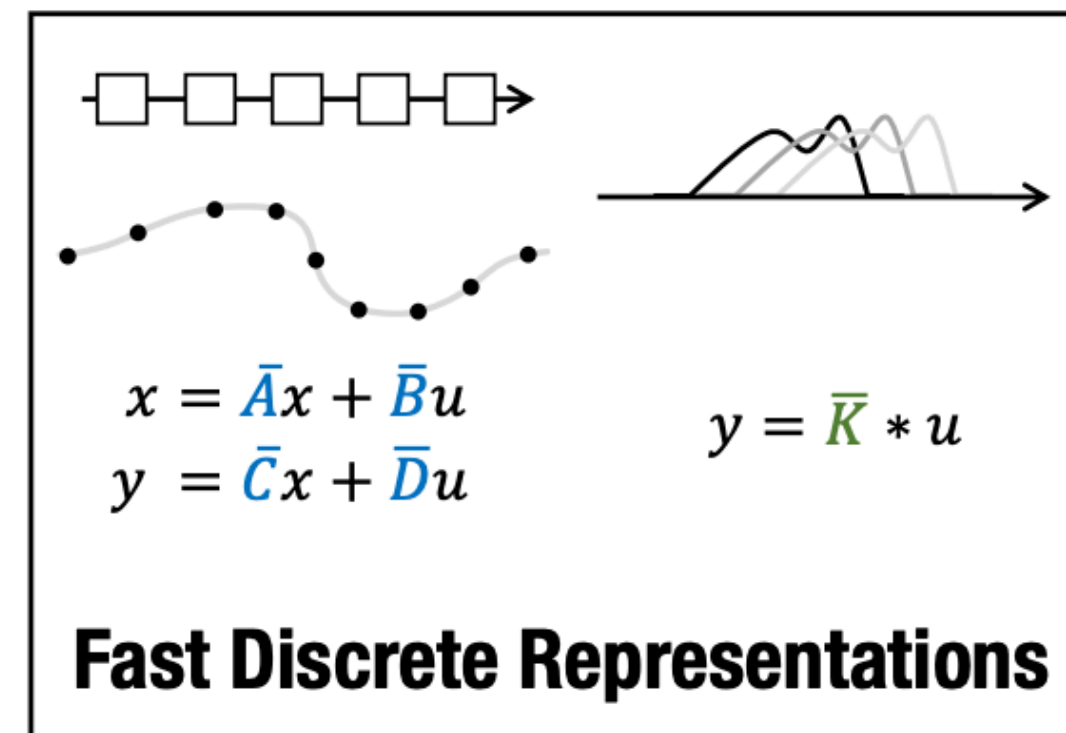
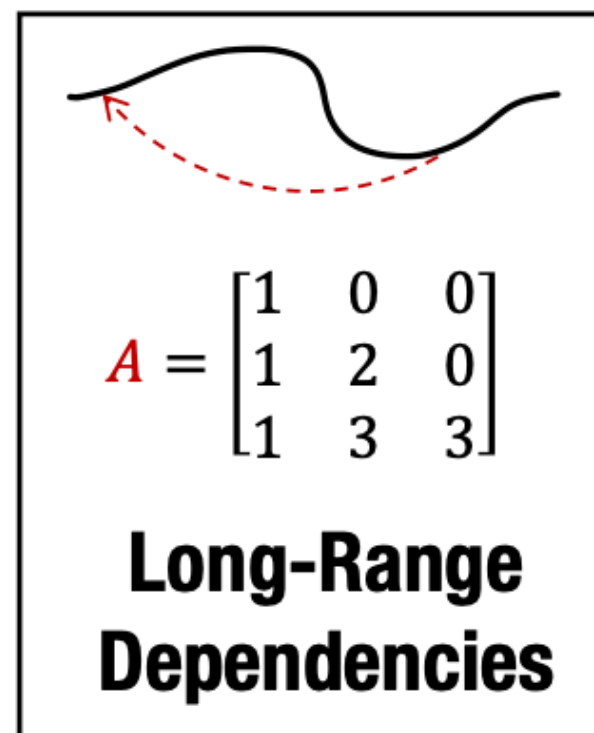
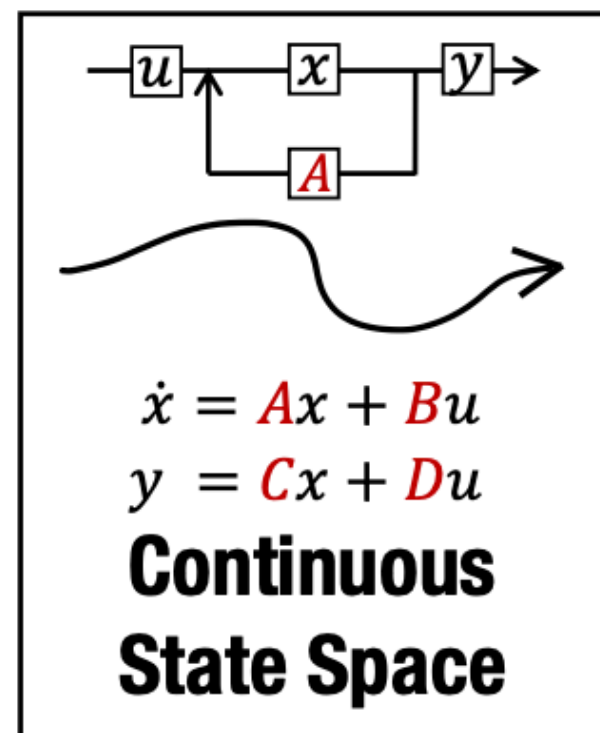
(b) 4th layer, left boundary detector



(c) 4th layer, right boundary detector

Units in the self-attention layers are interpretable

State Space Models



MODEL	LISTOPS	TEXT	RETRIEVAL	IMAGE	PATHFINDER	PATH-X	AVG
Transformer	36.37	64.27	57.46	42.44	71.40	✗	53.66
Reformer	<u>37.27</u>	56.10	53.40	38.07	68.50	✗	50.56
BigBird	36.05	64.02	59.29	40.83	74.87	✗	54.17
Linear Trans.	16.13	<u>65.90</u>	53.09	42.34	75.30	✗	50.46
Performer	18.01	65.40	53.82	42.77	77.05	✗	51.18
FNet	35.33	65.11	59.61	38.67	<u>77.80</u>	✗	54.42
Nyströmformer	37.15	65.52	<u>79.56</u>	41.58	70.94	✗	57.46
Luna-256	37.25	64.57	<u>79.29</u>	<u>47.38</u>	77.72	✗	<u>59.37</u>
S4	59.60	86.82	90.90	88.65	94.20	96.35	86.09

Nonlinear in depth, but *linear in time*, allows use of tricks like FFT for forward inference

Can handle extremely long context lengths - but don't yet compete with Transformers on language

A. Gu, K. Goel, C. Ré, ICLR (2022)

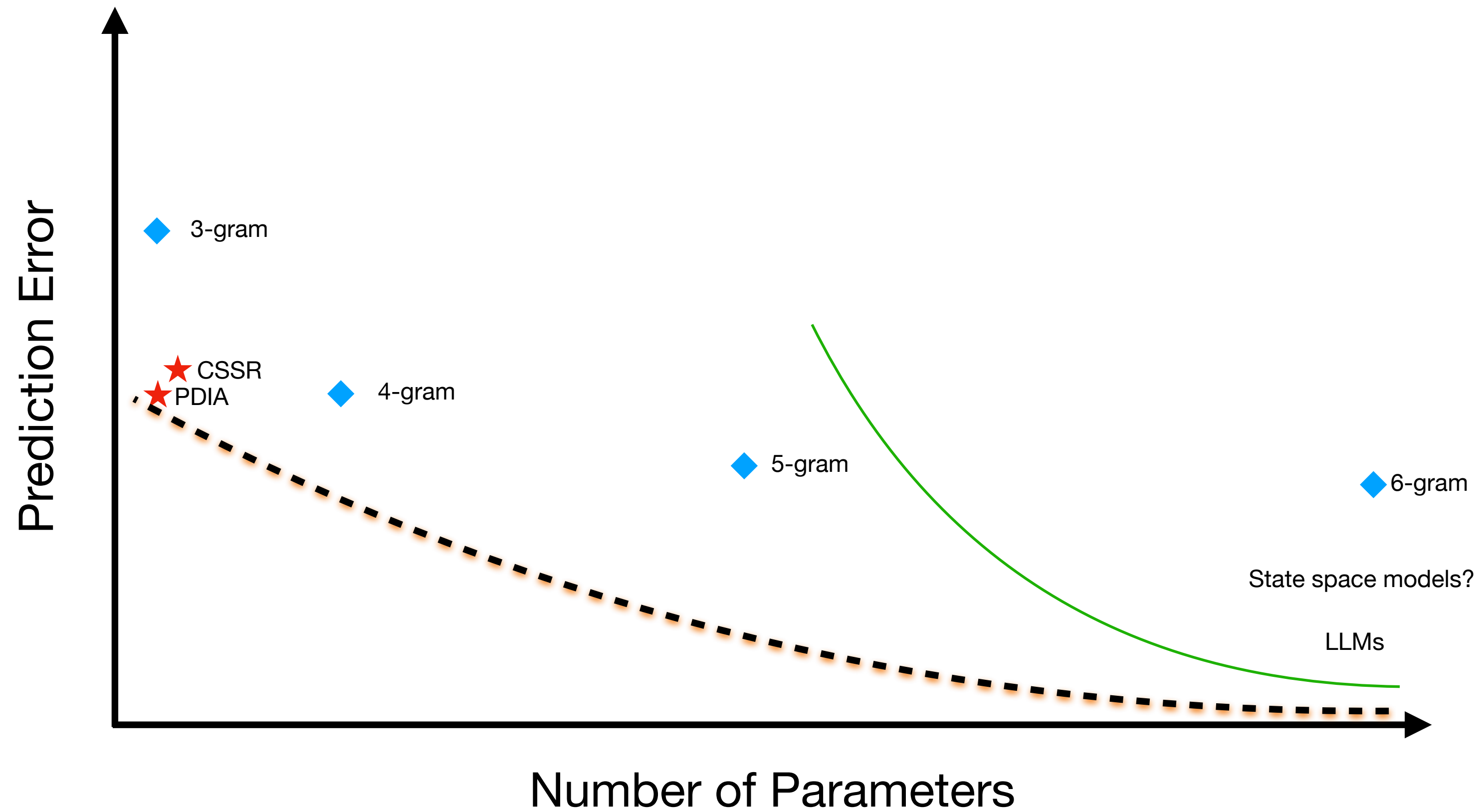
D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra, C. Ré, ICLR (2023)

J. T. H. Smith, A. Warrington, S. Linderman, ICLR (2023)

Why Small Models?

- Learn from *small data* (but, double descent...)
- Learn *interpretable* models (but, self-attention units...)
- Learn *quickly* and *efficiently*
- *Quantify* the amount of information in a time series - $H[\mathcal{S}]$ is an upper bound on predictive information

What next?



Still no models that cover the entire Pareto front!

Conclusions

- The problem of sequence prediction is largely a problem of understanding how to share information between contexts
- N-gram models suffer from the curse of dimensionality, while finite state machines are appealingly compact
- PDIA enables Bayesian nonparametric estimation of probabilistic state machines, but doesn't scale beyond small problems
- LLMs are outrageously successful at modelling complex time series, but are not parsimonious with their parameters
- Is it possible to build models that are efficient with simple time series, but can still scale?

 pfau@google.com

 @pfau

Thank you!