The Serial Scaling Hypothesis

Yuxi Liu*

Konpat*

Kananart

Yutong Bai

*Equal contributions

The Plan

30-min Presentation

30-min QA
(but feel free to ask burning questions anytime 🔥)

What is a serial problem?

“9 Women cannot make a baby in a month”

↓

Human development is a serial problem!

For us ML people:

~~“9 Women cannot make a baby in a month”~~

Reasoning

Decision making (RL)

Simulating dynamic systems

Physics

Video prediction

The Serial Scaling Hypothesis (SSH):

These problems need to scale more serial compute,
Not just parallel ones.

*sound intuitive enough? But we didn’t quite follow this intuition.

We didn’t quite follow the **serial** intuition

2017: We ditched RNN (serial) => Transformers (parallel)

2021: Scaling law doesn’t make distinction serial/parallel

2024: Test-time scaling doesn’t make distinction serial/parallel

Now: We use Diffusion models to do visual reasoning

What does Serial Scaling Hypothesis do?

Explain past successes

Connect to complexity theory

Connect to practice

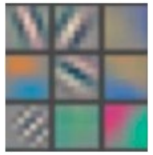
Finally, implications of this..

Explains past successes

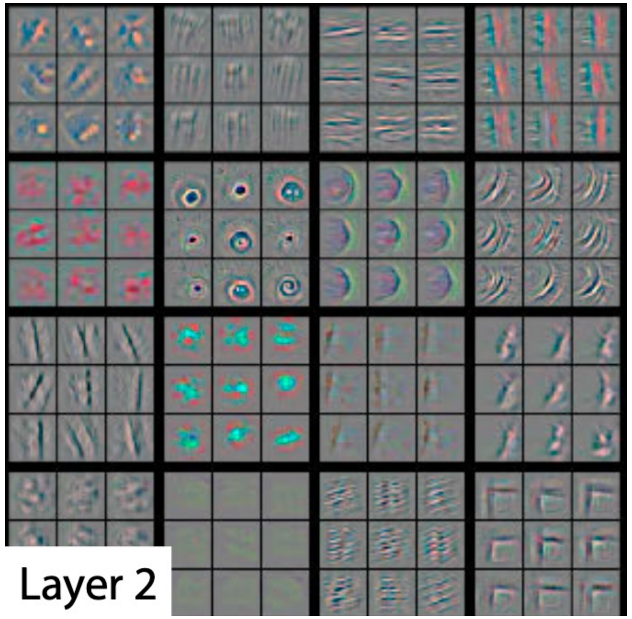
Past success:

Deep learning is powerful because it is "deep"

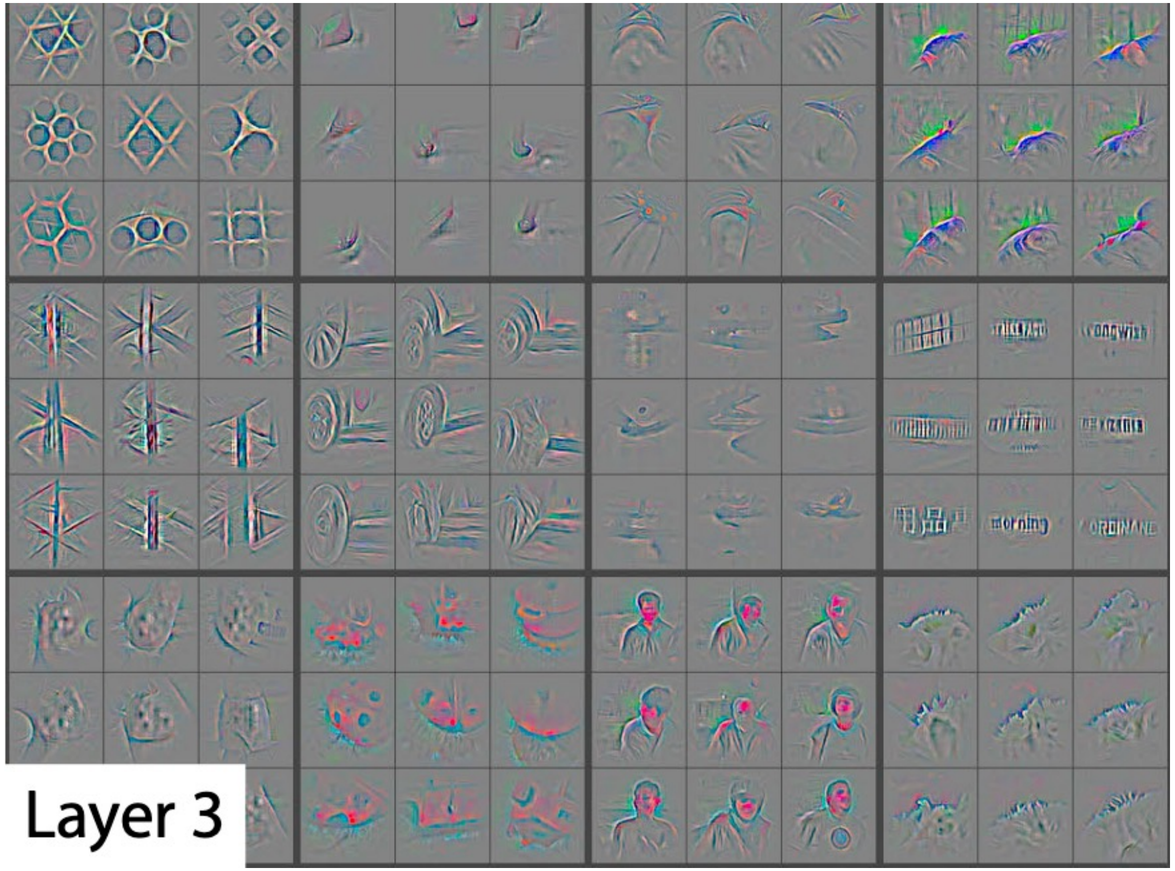
Zeiler & Fergus 2014



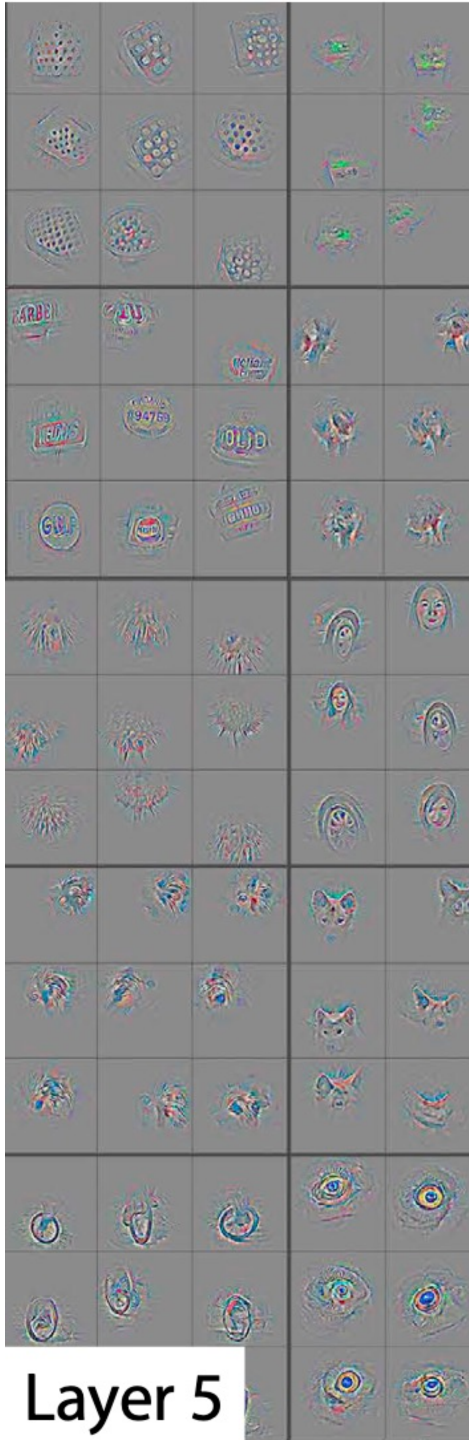
Layer 1



Layer 2



Layer 3



Layer 5

*Shallow learning: Imagine everything in the first layer.

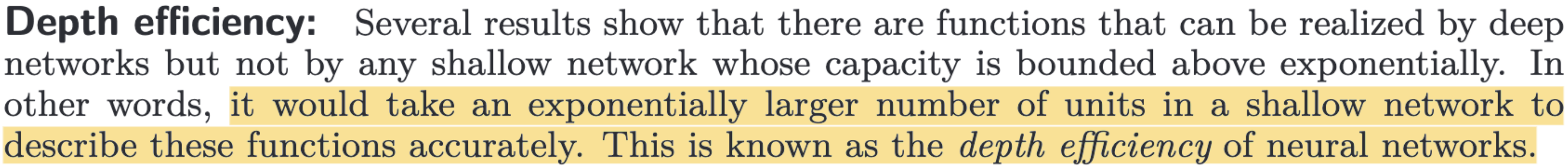
Past successes

Chain-of-Thought improves LLMs because it's "deeper"

Li 2024, Merrill 2024

Exponential depth-width trade-off

Prince 2023 (and many many others)

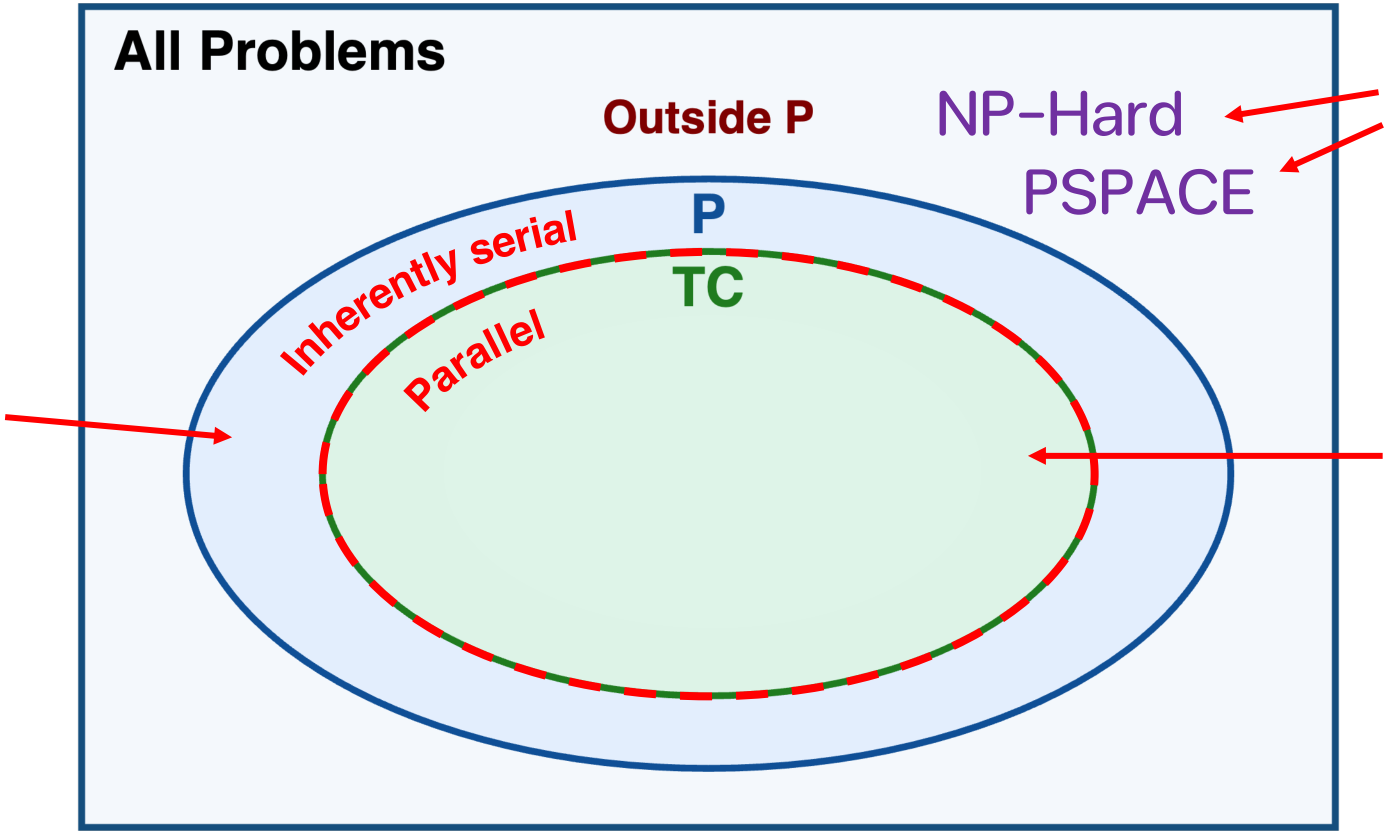


Depth efficiency: Several results show that there are functions that can be realized by deep networks but not by any shallow network whose capacity is bounded above exponentially. In other words, it would take an exponentially larger number of units in a shallow network to describe these functions accurately. This is known as the *depth efficiency* of neural networks.

Connects to complexity theory

Some problems are not efficiently parallelizable.
They need increasing serial compute.

P-complete not known how to parallelize



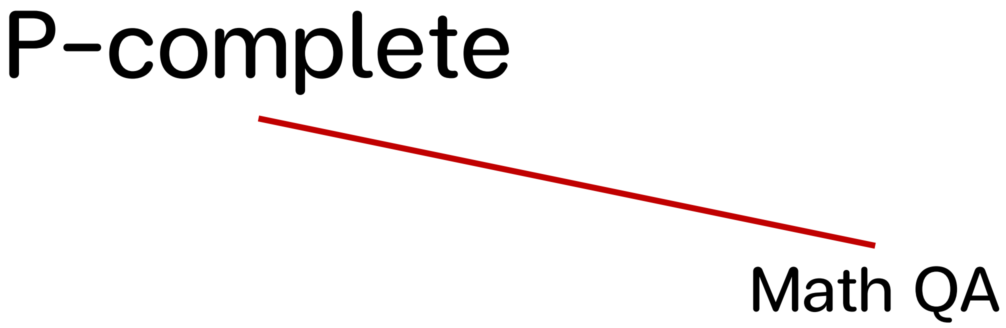
We don't care about these

TC = efficiently parallelizable

*Greenlaw 1995 (assume P != TC)

*Bottleneck is not "serial". We don't know how to solve such hard problems.

Inherently Serial Problems



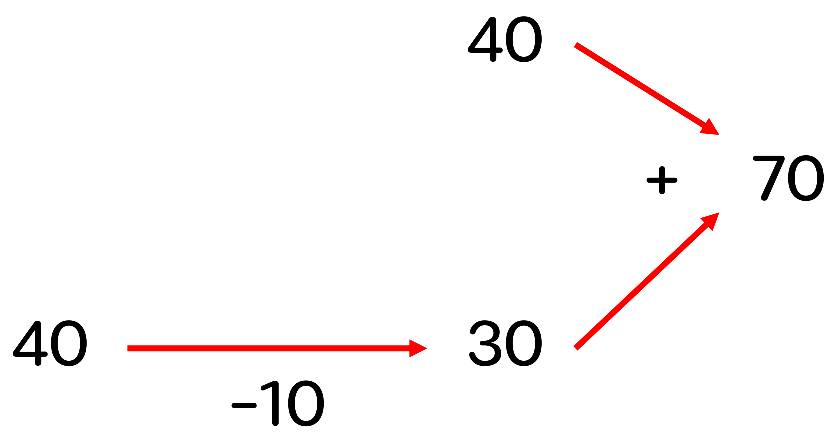
P-complete

Math QA

Math QA is serial…

GSM8K: James spends 40 years teaching.
His partner has been teaching for 10 years less.
How long is their combined experience

Graph:

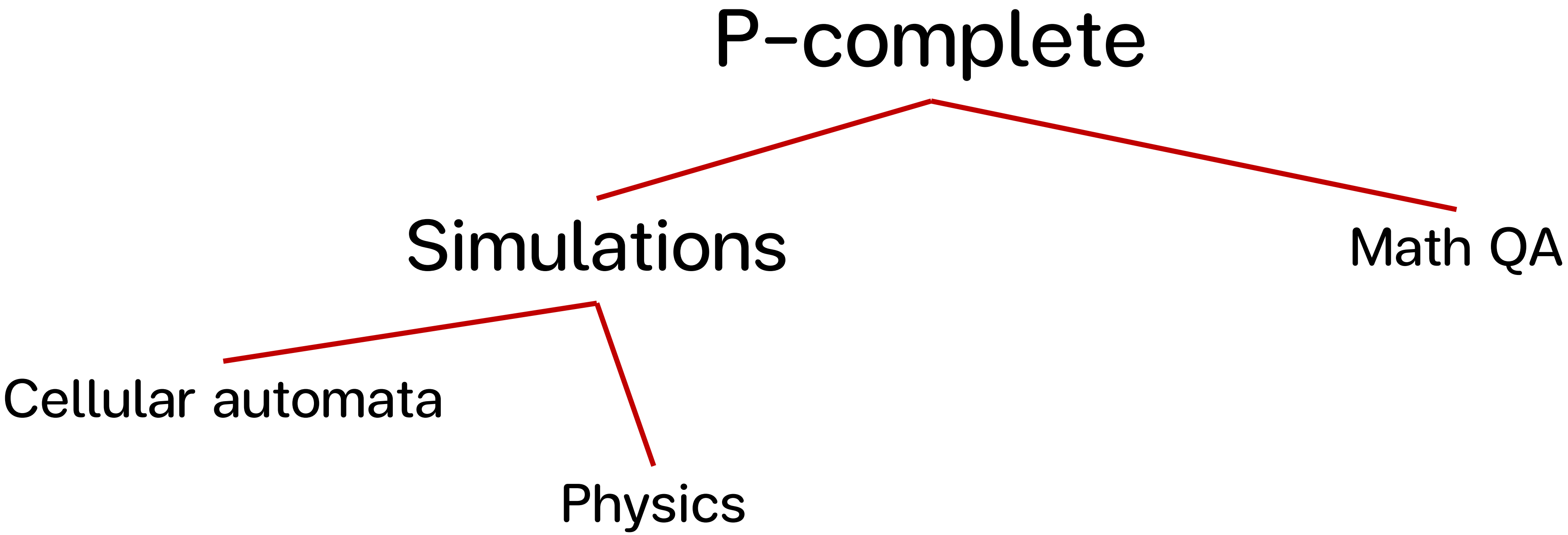


Arithmetic CVP (P-complete)

*Greenlaw 1995

*As serial as the depth of the graph

Inherently Serial Problems



P-complete

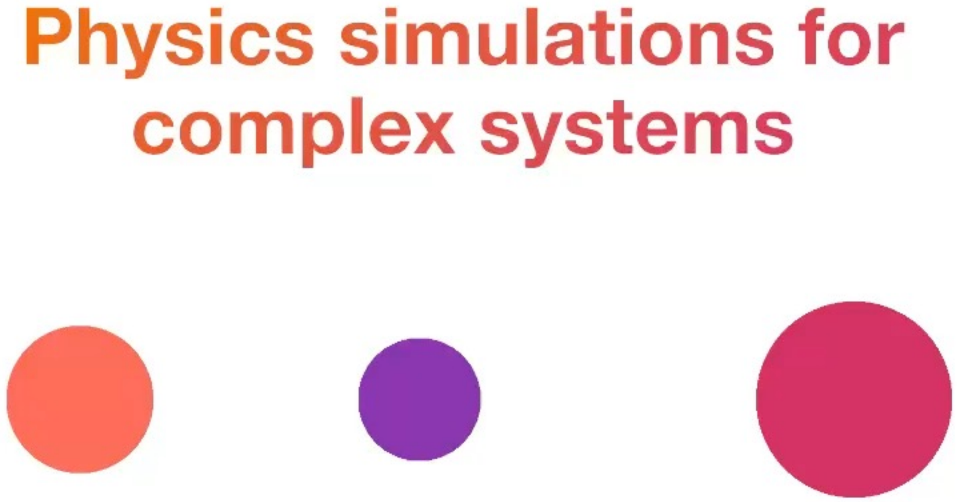
Simulations

Math QA

Cellular automata

Physics

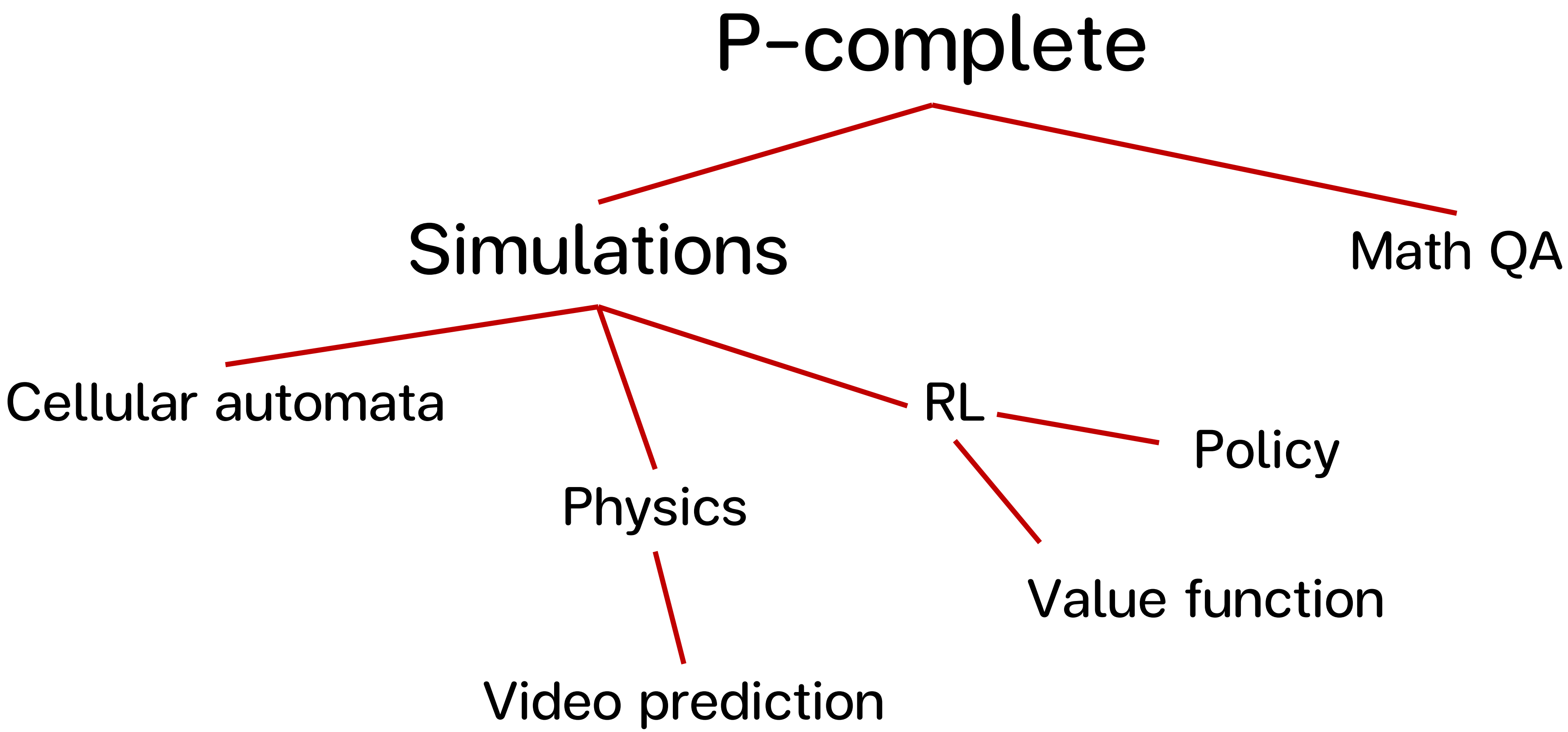
Simulations are inherently serial



No shortcut solution for row N

No shortcut solution for frame T

Inherently Serial Problems



P-complete

Simulations

Math QA

Cellular automata

Physics

RL

Policy

Value function

Video prediction

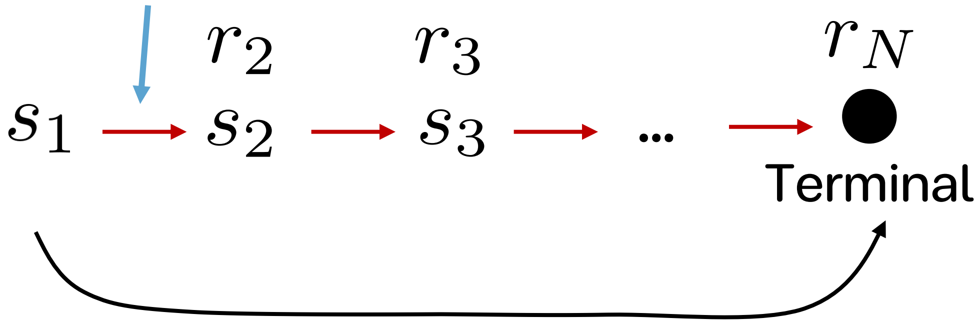
*they are serial in general case

RL's Value function is serial problem

Value function $R = \sum^{N} r_i$

Serial problem!

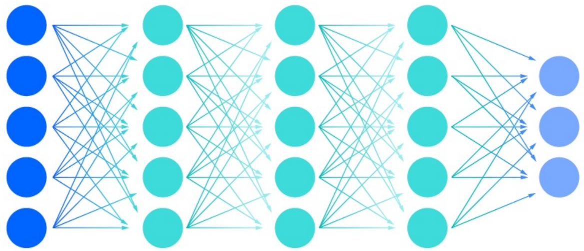
Transition = policy + environment



No shortcut to state N

Model-free RL
V(s)

s_1 →

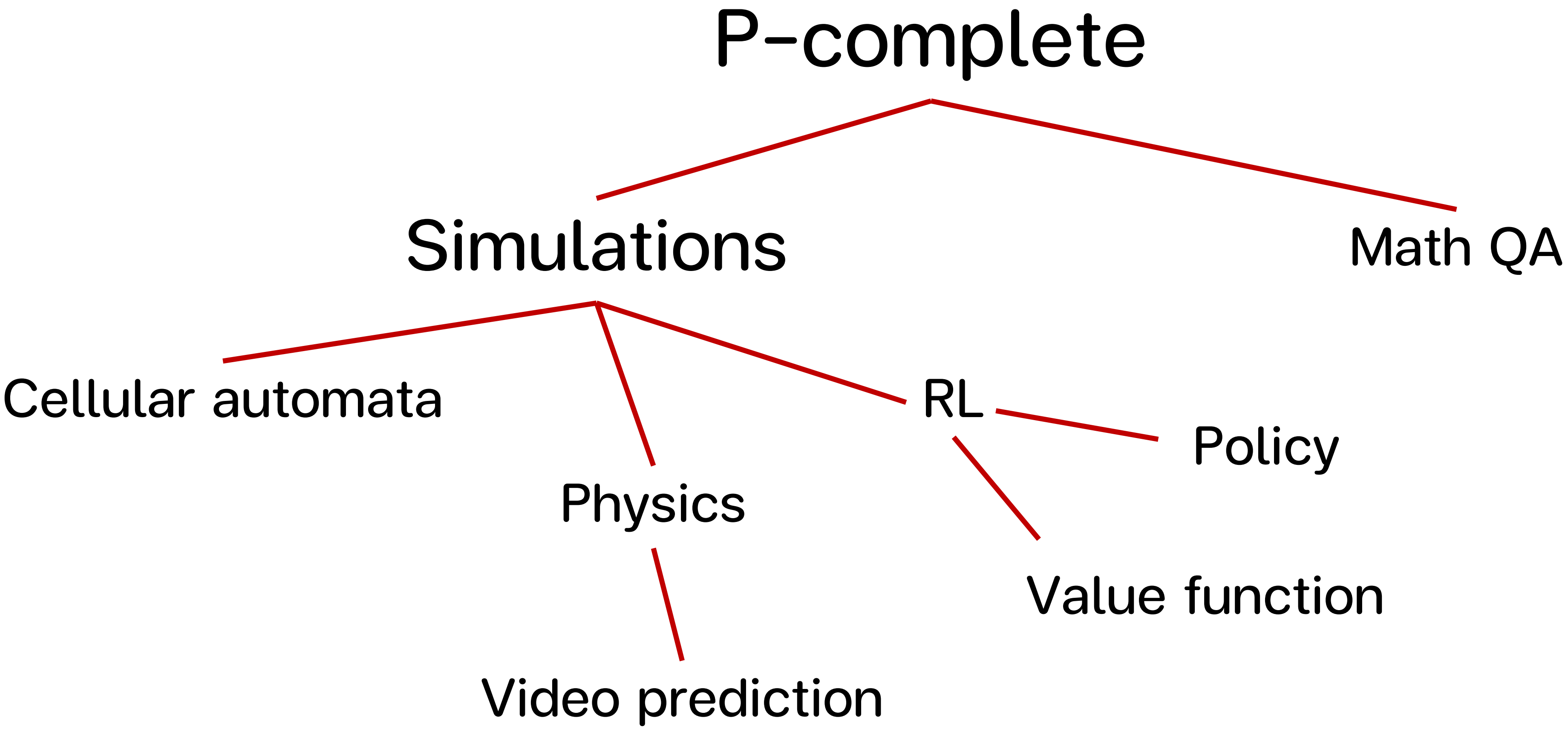


→ $R \neq \sum^{N} r_i$

Shallow neural network
*Biased est.

*Unbiased R is required for optimal policy!

Inherently Serial Problems



*more details in the paper

Connects to practice

Practice

LLMs struggle with math/reasoning problems

LLM solves arithmetic with
“bag of heuristics” (Nikankin 2024)

CoT improves math/reasoning
(Kojima 2022)

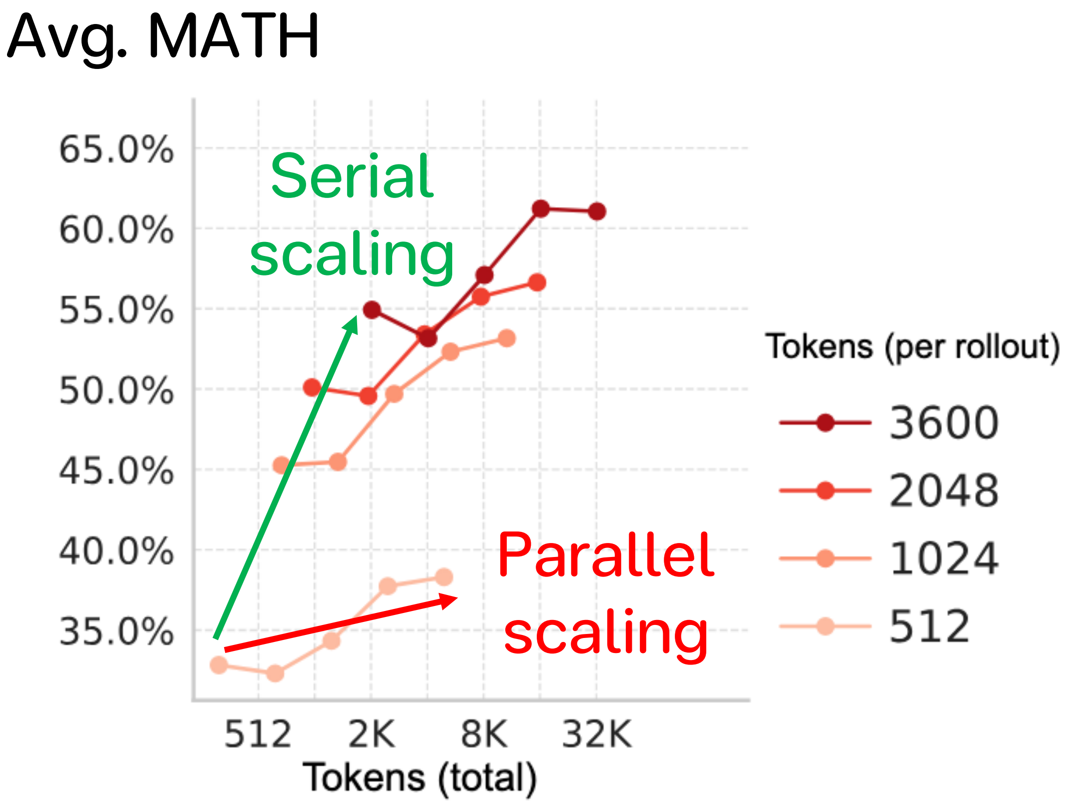
Theory

Transformer has limited serial compute
(Merrill 2023)

This includes Mamba (Merrill 2024)
and other SSMs.

CoT increases serial compute of
Transformers (Li 2024, Merrill 2024)

Math QA scales better with serial…

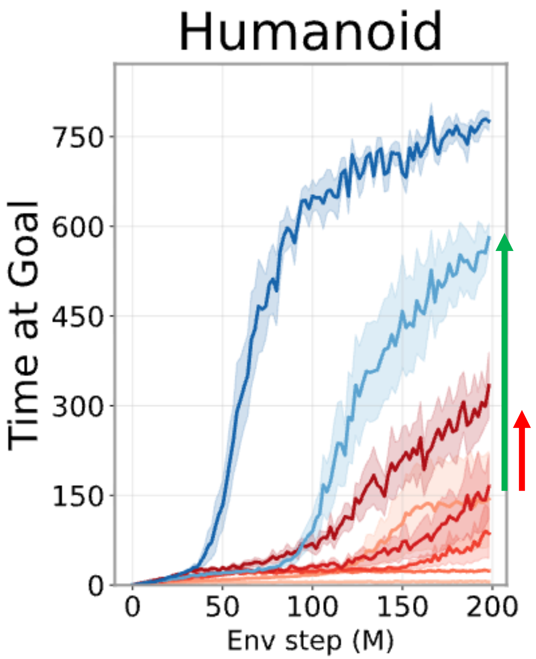


*Aggarwal & Welleck 2025

Practice

Model-based RL > Model-free in Go
(Silver 2016, 2017)

Deeper > Wider value & policy networks
(Kevin 2025)



2x depth

16x width

256 width

Theory

Model-based RL is more serial
than model-free RL

Deeper network is more serial

Practice

Your

Theory

?
Let’s think together during the QA.

In terms of models…

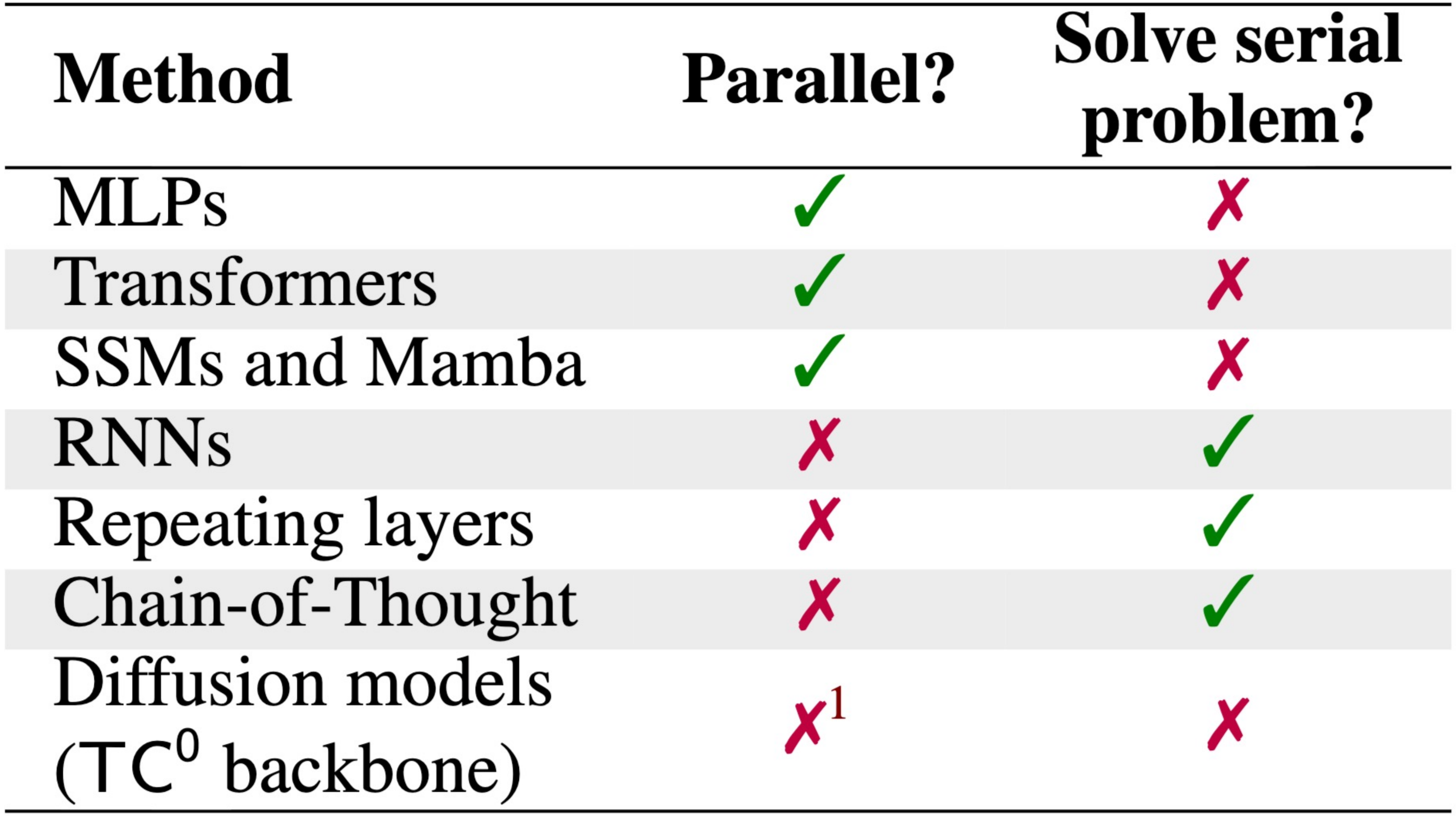
Parallel

Serial

Feed forward

Recurrent

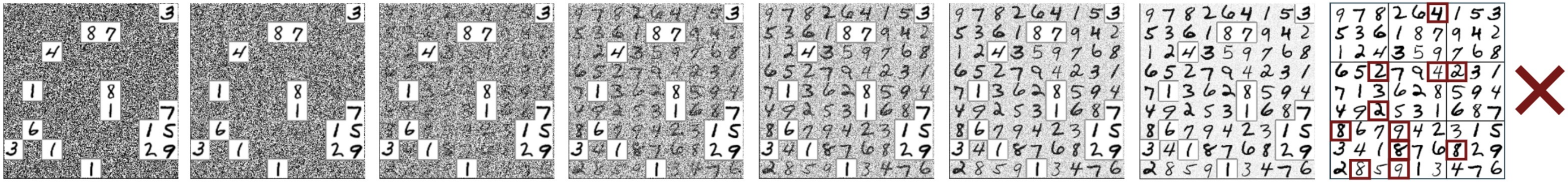
New finding: **Parallel!**



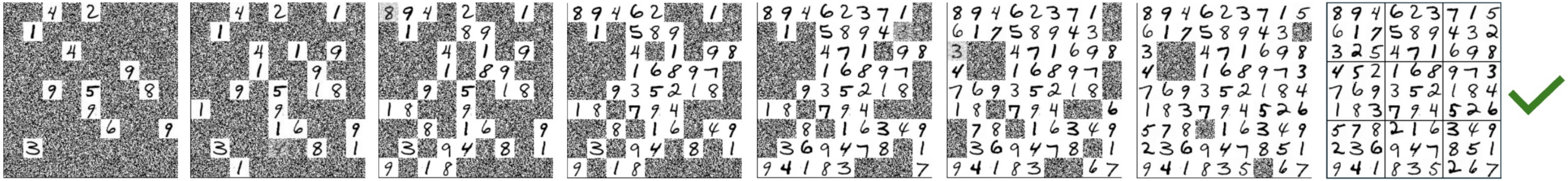
Method	Parallel?	Solve serial problem?
MLPs	✓	✗
Transformers	✓	✗
SSMs and Mamba	✓	✗
RNNs	✗	✓
Repeating layers	✗	✓
Chain-of-Thought	✗	✓
Diffusion models (TC^0 backbone)	✗[1]	✗

*Merrill 2022, Chiang 2024, Chen 2024, Merrill 2024

Solving sudoku with diffusion



Solving sudoku with autoregressive



*Wewer 2025

Practice

Diffusion models don’t scale well with more steps

Image generation (Karras 2022)

Depth estimation (Ravishankar 2024)

Language modeling (Austin 2021)

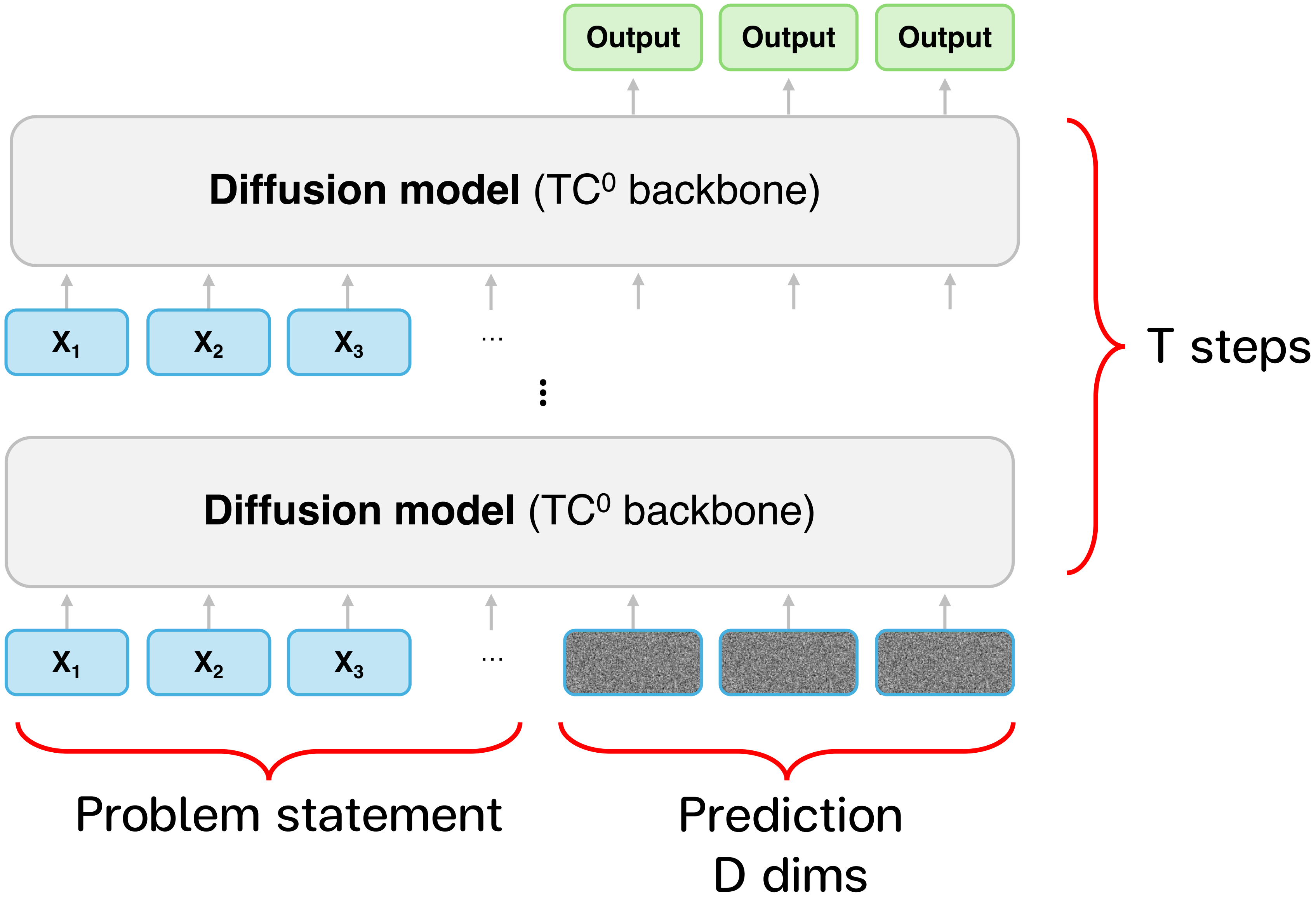
Theory

New:

If a problem can be solved by a diffusion model (with a TC^0 backbone) with high probability with infinite diffusion steps, then the problem itself is in the parallelizable class TC^0.

The idea of the proof is not complex

Diffusion model setup



*D can be constant

Diffusion proof steps

How fast does diffusion approach solution to problem (of size N, output D dims)

T-step distribution p_T

inf-step distribution p_∞

$$TV(p_T, p_\infty) = O(D/T)$$

(Li and Yan 2024)

Close enough to solve* ϵ

$$\epsilon = O(D/T) \qquad T = O(D/\epsilon)$$

Diffusion solves problem at the rate independent of N
(Any problem can be solved in constant sampling steps)

What diffusion solves is not a serial problem

Intuition: diffusion score function is “smooth” (converge in few steps…)

*Assume diffusion models any score function. *”Close enough” is explained in the paper.

That’s the gist of the paper…
Now,

Implications

Implication of SSH

We still need higher clock CPUs!

Because serial compute cannot be substituted by parallel compute!

Need new serial models and how to train them

Need recurrence in models.
How to deal with training instability?

Might explain data hungriness:

Insufficient depth,
Exponential width (model size),
Exponential data needed?

Last resort: “If you cannot solve the proposed problem, try to solve first some related problem” (Polya 1957)

Approximation.
Change: Truncated RL (Park 2025, Sutton 2018).
Inspiration from math, factorization is hard, primality is easier.

Thank you! 🙏
QA