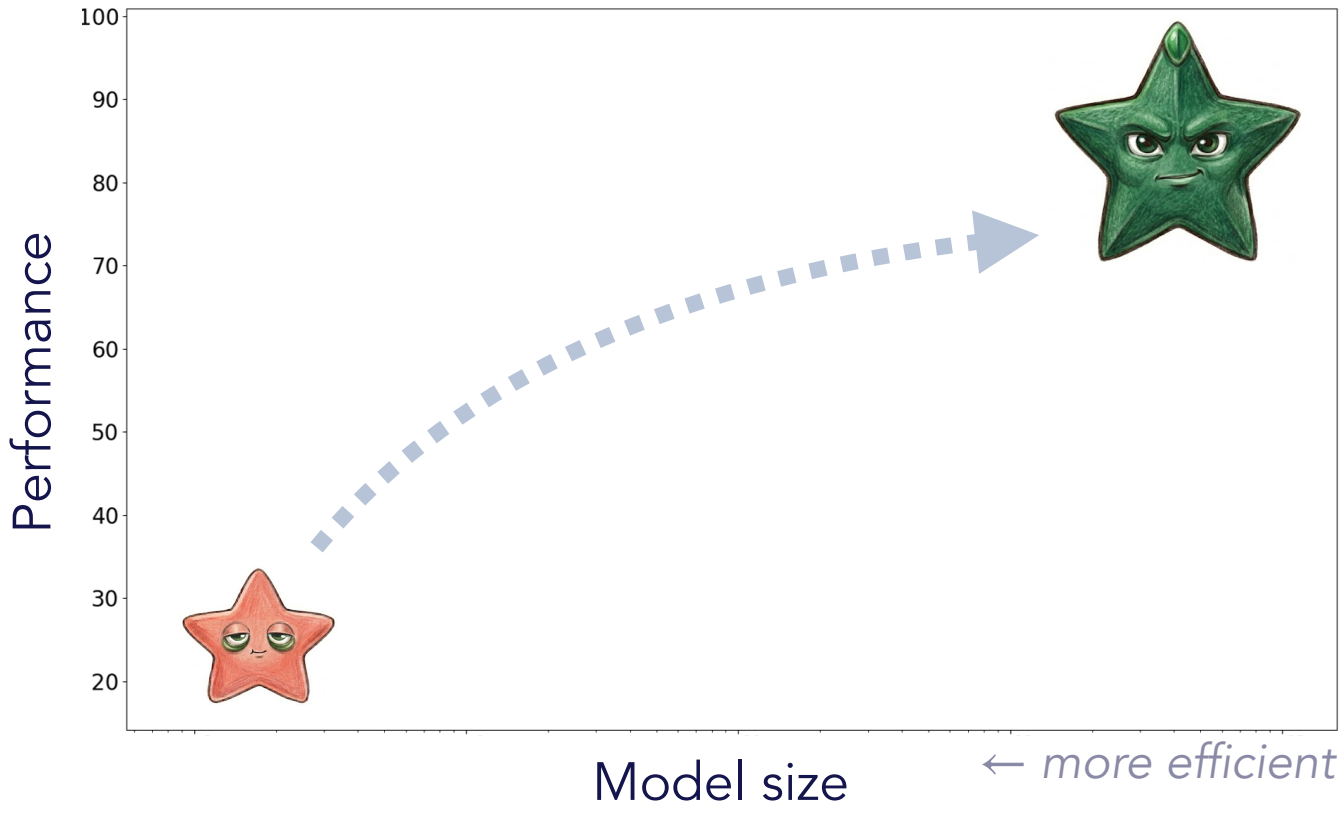
# Understanding and Realizing the Potential of Smaller Models

Bingbin Liu

Kempner Institute, Harvard University

# Model size limits capability?

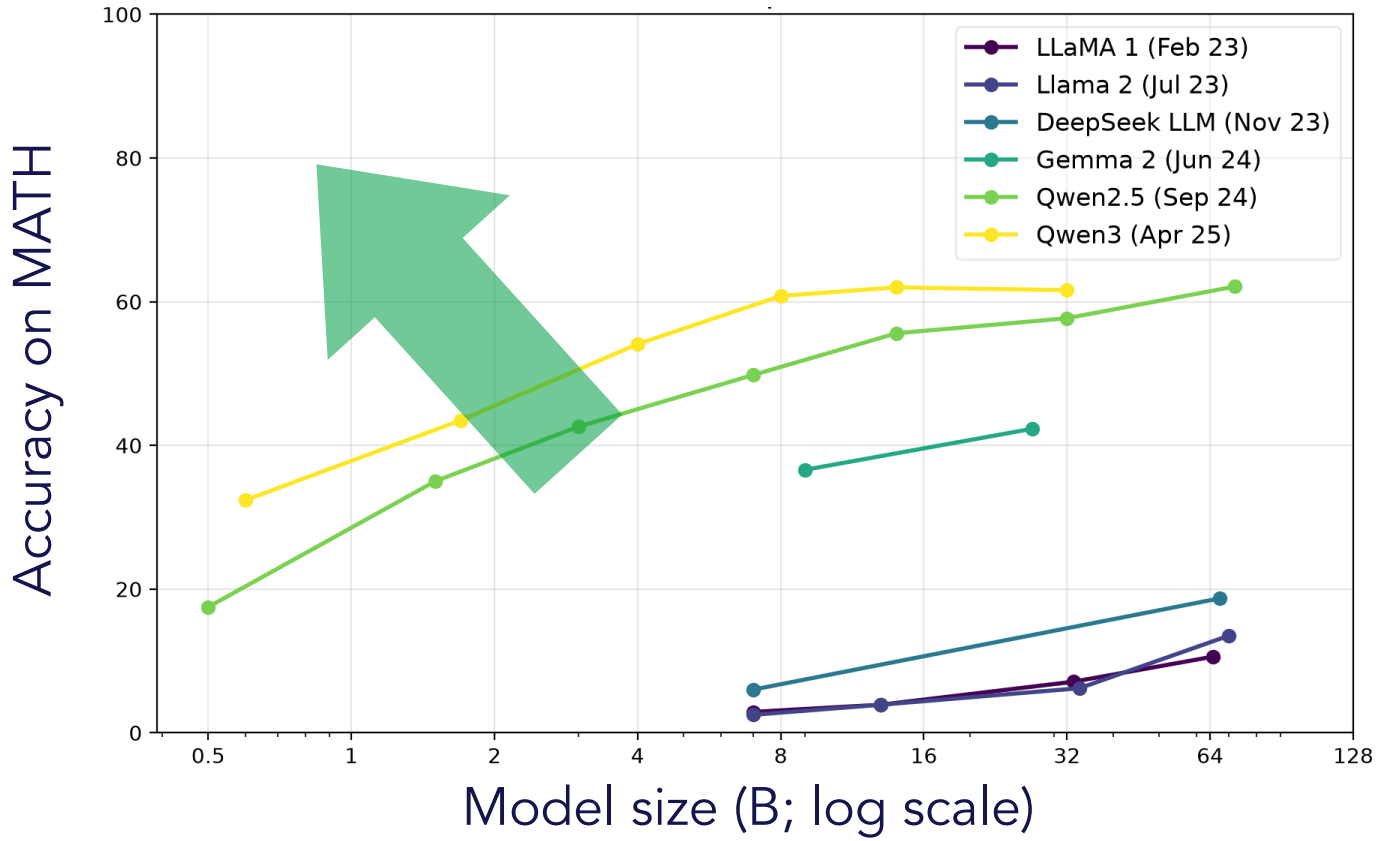


# Lifting the scaling curve

*(size-vs-capability)*

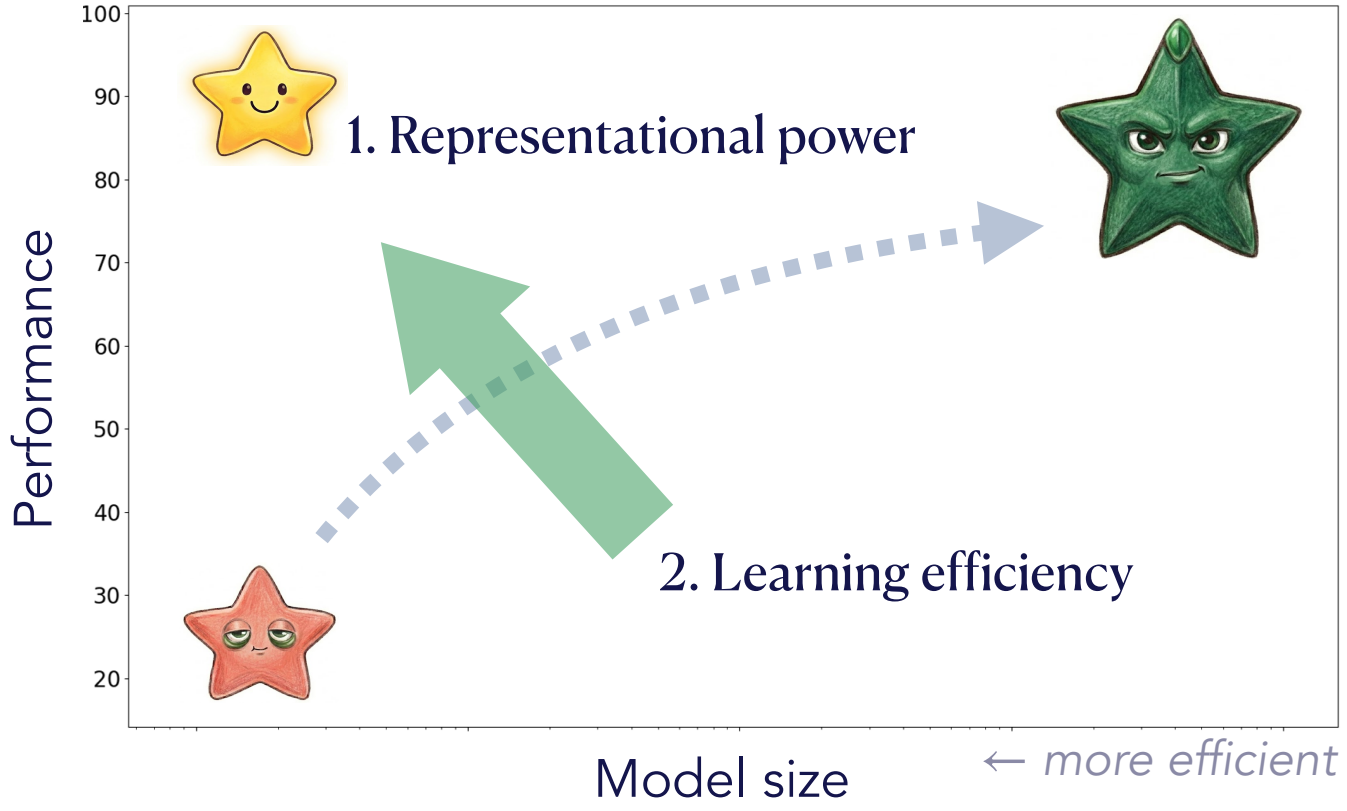
Understanding
(fundamental limits of efficient reasoning?)

Cost
(accessibility and applications)




MATH 4-shot on base

# Lifting the scaling curve — *how much?*

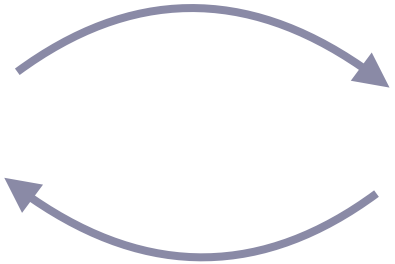


# Understanding via abstractions (sandboxes)

**Reality**
(messy, complex)

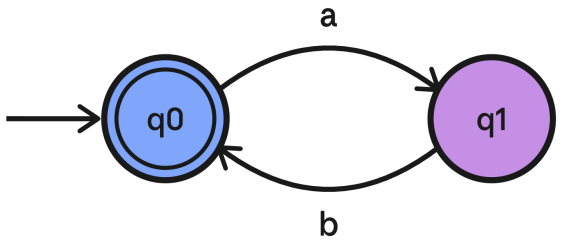



understanding



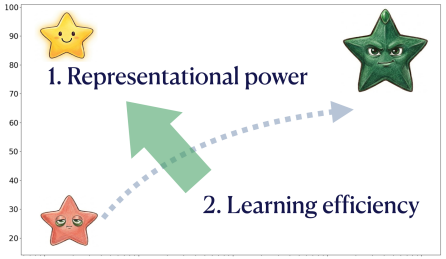
practical improvements

**Abstraction**
(clean, simple)



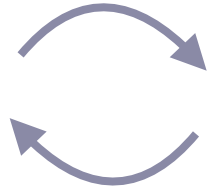

*Smaller scale benefits more*
*from algorithms + ideas!*

# Lifting the scaling curve

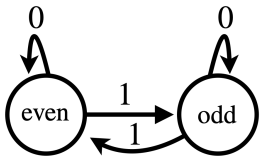



## Part 1: Representational power

Parallel compute for sequential reasoning



automata



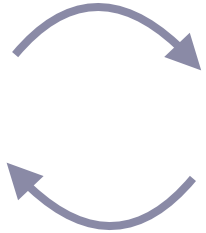

*diagnosing models ← shallow constructions*

LAGKZ: Transformers Learn Shortcuts to Automata. *ICLR 23 (Oral)*.

LAGKZ: Exposing Attention Glitches with Flip-Flop Language Modeling. *NeurIPS 23 (Spotlight)*.

## Part 2: Learning efficiency

Improving distillation



sparse parity

$x = 1\ -1\ -1\ 1\ -1\ 1$

*how to choose ← why teacher choice matters*

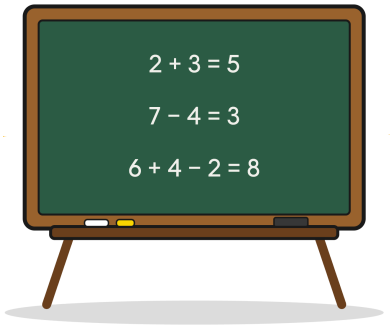
P*L*MRG: Progressive distillation induces an implicit curriculum. *ICLR 25 (Oral)*.

PLMKG: In Good GRACEs: Principled Teacher Selection for Knowledge Distillation. *ICLR 26*.

# Part 1: Compact reasoning solutions

Sequential reasoning

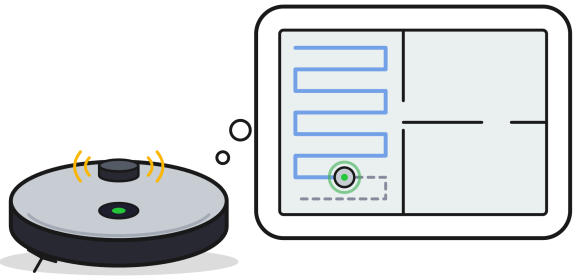
?




Math



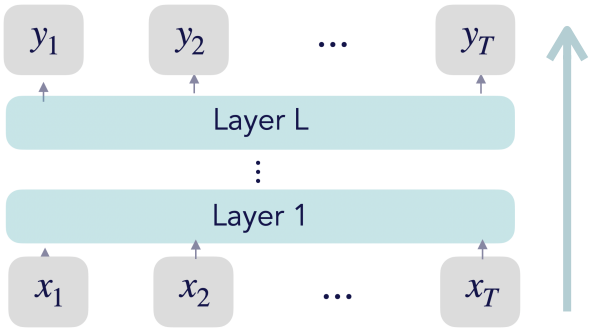

Coding



State tracking

Parallel computation

(*Transformer*)




*Role of **depth**?*

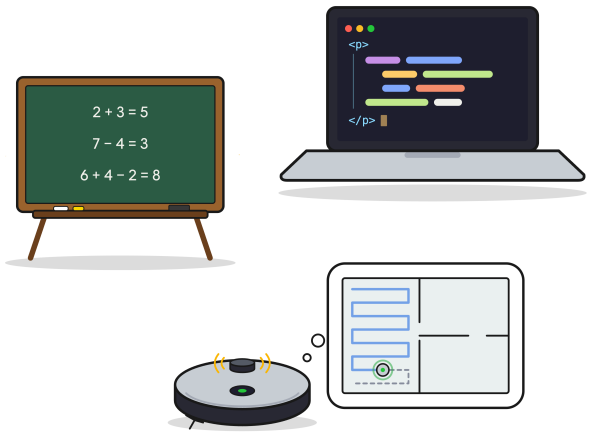
*for computation of one forward pass*

# Part 1: Compact reasoning solutions

*computation as state transition*

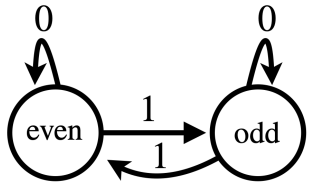


## Sequential reasoning




diagnosing learned models

## Automata (DFA)




Shallow solutions exist.

# Sequential reasoning via automata

$\mathcal{A} := (Q, \Sigma, \delta)$

states inputs transitions

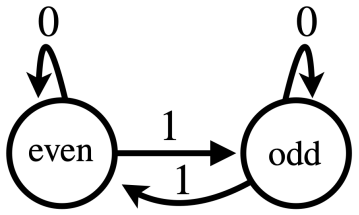
$q_t = \delta(\sigma_t, q_{t-1})$

( $Q$ is finite )

parity counter

$Q = \{\text{even}, \text{odd}\}$

$\Sigma = \{0, 1\}$

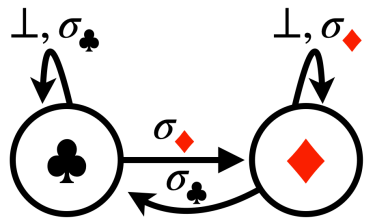



1-bit memory unit

$Q = \{\clubsuit, \diamondsuit\}$

$\Sigma = \{\sigma_\clubsuit, \sigma_\diamondsuit, \perp\}$

(no-op)

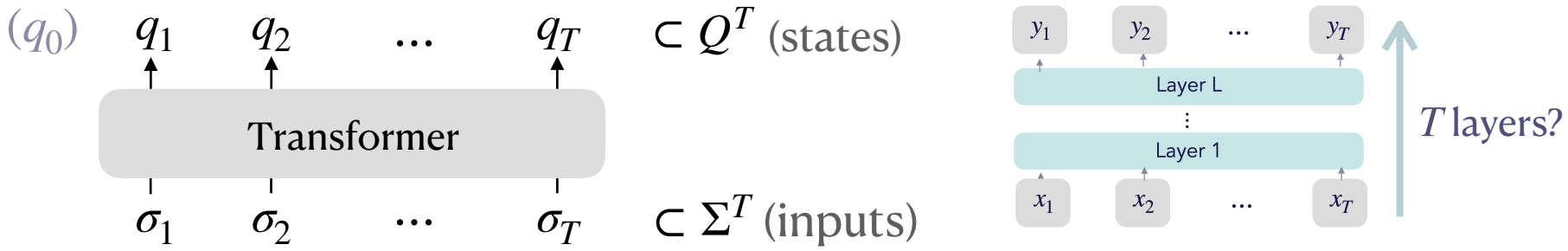


# Sequential reasoning via automata

$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

Reasoning: simulating $\mathscr{A}$: learn a seq2seq function with length $T$.

(trivial with $T$ steps)




1. Any $\mathscr{A}$: $O(\log T)$ layers

*Benefit of looping in models*

2. Special $\mathscr{A}$: $O(|Q|^2 \log |Q|)$ layers

*Reliability challenges and improvements*

# Simulating $T$ transitions in $\mathscr{A}$

$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

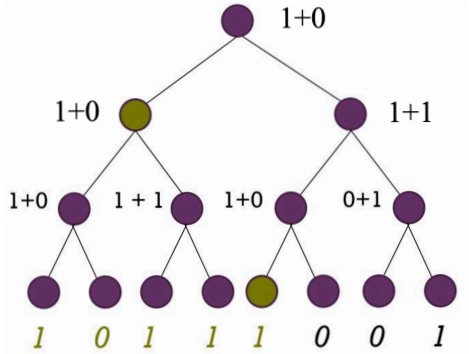
1. $O(\log T)$ layers for any $\mathscr{A}$: divide-and-conquer.

**function composition**

$$q_T = \delta(\sigma_T, \delta(\cdots, \delta(\sigma_2, \delta(\sigma_1, q_0)))) = (\underbrace{\delta(\sigma_T, \cdot)}_{\mathbb{R}^{|Q|\times|Q|}} \circ \cdots \circ \delta(\sigma_1, \cdot))(q_0)\,.$$

**associativity**

$$f_1 \circ f_2 \circ f_3 \circ f_4 = (f_1 \circ f_2) \circ (f_3 \circ f_4)$$




$\log T$

- aka. "associative scan" in **prefix sum** [Blelloch 93], subquadratic models e.g. **Mamba** [Gu & Dao 23].

# Simulating $T$ transitions in $\mathscr{A}$

$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

1. $O(\log T)$ layers for any $\mathscr{A}$: divide-and-conquer.

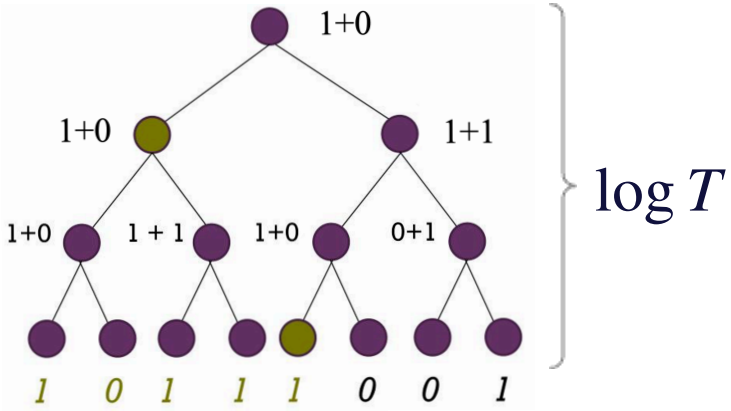
- *No shallower* in the general case: $\Omega(\log T)$

  *assuming* $TC^0 \neq NC^1$. [Merrill & Sabhawal 22]

- More layers can increase expressivity.

  $O(1) \rightarrow O(\log T)$ depth: provably more powerful.

  e.g. looping! [Merrill & Sabhawal 25]



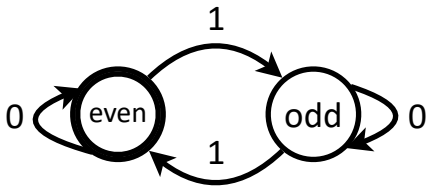
# Simulating $T$ transitions in $\mathscr{A}$

$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

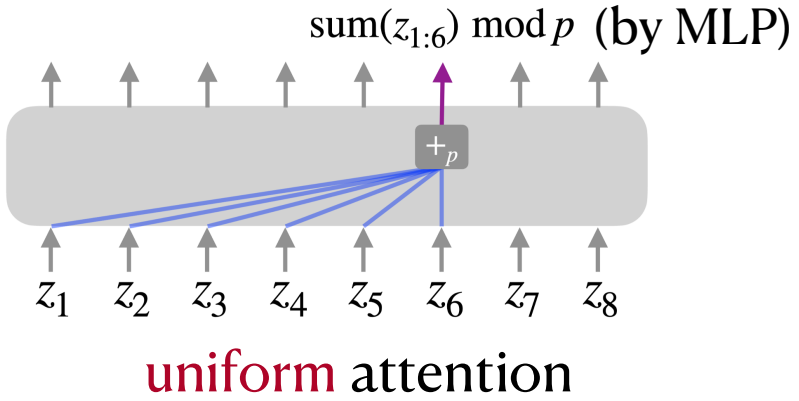
2. $O(|Q|^2 \log |Q|)$ layers for special (solvable) $\mathscr{A}$.

- Idea: certain compositions are highly parallelizable.
  $O(1)$ layers if ***commutative***

e.g. parity:




$$q_T = \left(\sum_{t \in [T]} x_i\right) \mod 2$$



# Simulating $T$ transitions in $\mathscr{A}$

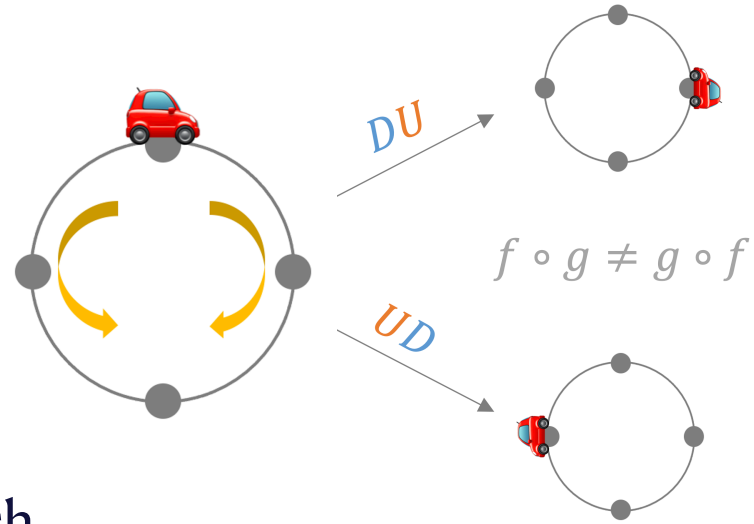
$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

2. $O(|Q|^2 \log |Q|)$ layers for special (solvable) $\mathscr{A}$.

- **Krohn-Rhodes** decomposition: factorize into parallelizable "base cases".

$Q = \{$ 🚗 , 🚗 $\} \times \{0,1,2,3\}$, $\Sigma = \{D$ (drive), $U$ (U-turn)$\}$.

$q_0 = \{$🚗$,0\}$, $\sigma_{1:T} = DDDUDDUUD \rightarrow q_T$?




1. Direction: parity of $U$.
2. Position: signed sum of $D$ (mod 4).

$O(1)$ layer each.

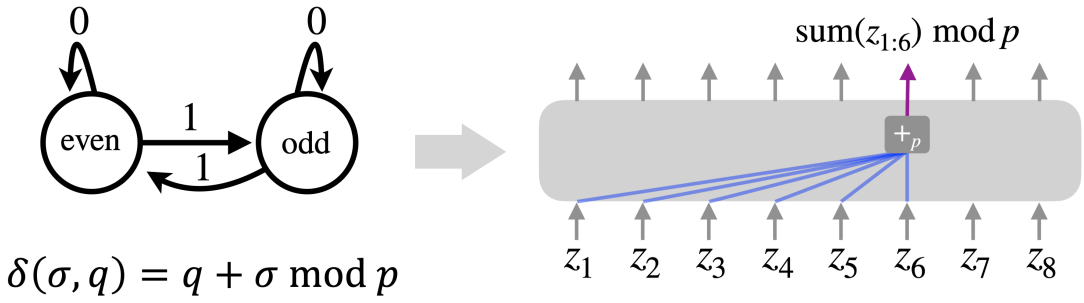
# Simulating $T$ transitions in $\mathscr{A}$

$$\mathscr{A} := (Q, \Sigma, \delta)$$
$$q_t = \delta(\sigma_t, q_{t-1})$$

2. $O(|Q|^2 \log |Q|)$ layers for solvable $\mathscr{A}$.

- **Krohn-Rhodes** decomposition: factorize into parallelizable "base cases".

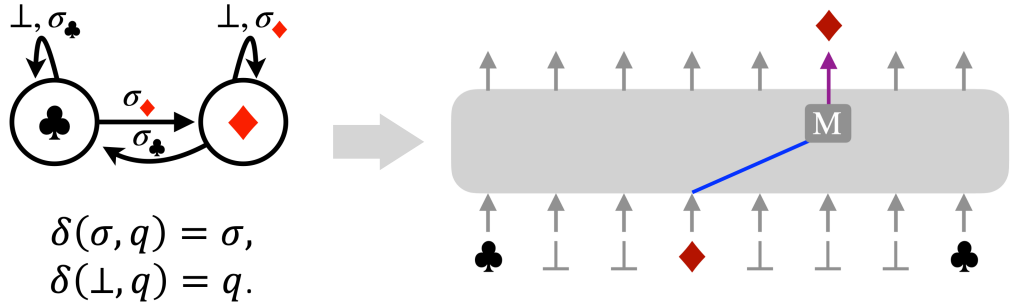
***2 cases**, each representable by 1 Transformer layer.*




mod(-$p$) counter

uniform attention

e.g. count, sum, average



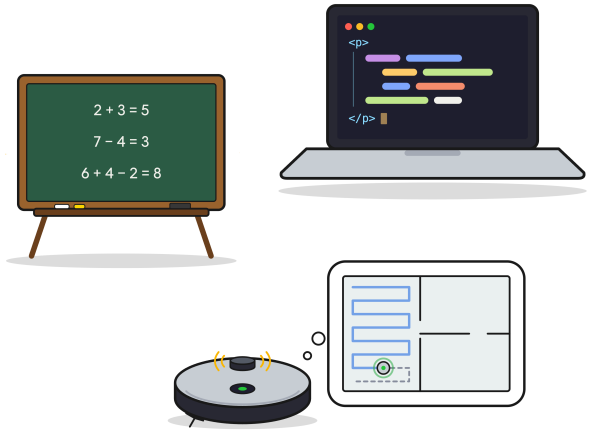

memory unit

1-sparse attention

e.g. selection, long-range, induction head.

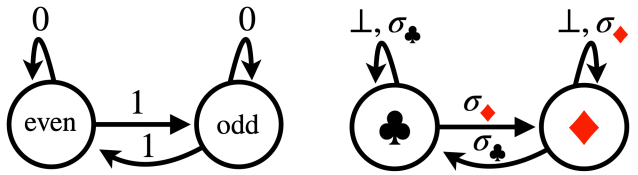
# Part 1: Compact reasoning solutions

## Sequential reasoning




Examining *learned* solutions

## Automata (DFA)



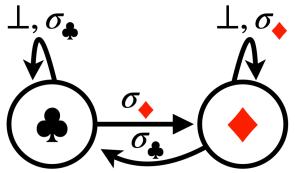

Any $\mathscr{A}$: $O(\log T)$ depth

Solvable $\mathscr{A}$: $O(|Q|^2 \log |Q|)$ depth

*$O(1)$-depth for gridworld from interpreting learned attention.*

# How well can models *learn* a "base case"?

Memory unit



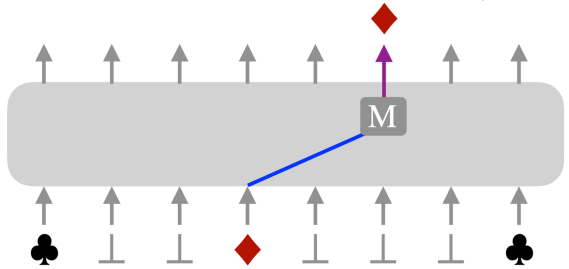

- Irrelevant contexts [Shi et al. 23, Yang et al. 25]

  Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?

- Role spoofing attack.



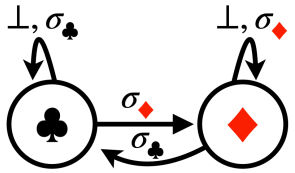

*Recall: representable with 1 Transformer layer*




(sparse attention)

# How well can models *learn* a "base case"?

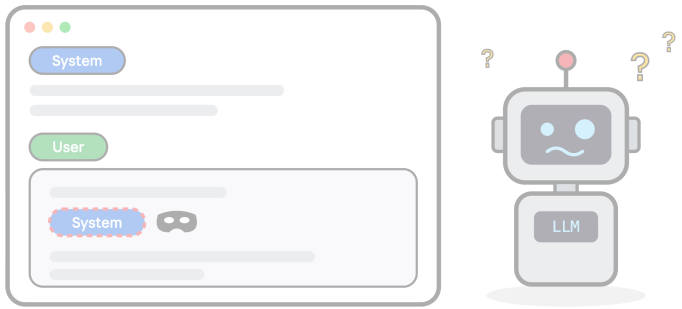
Memory unit



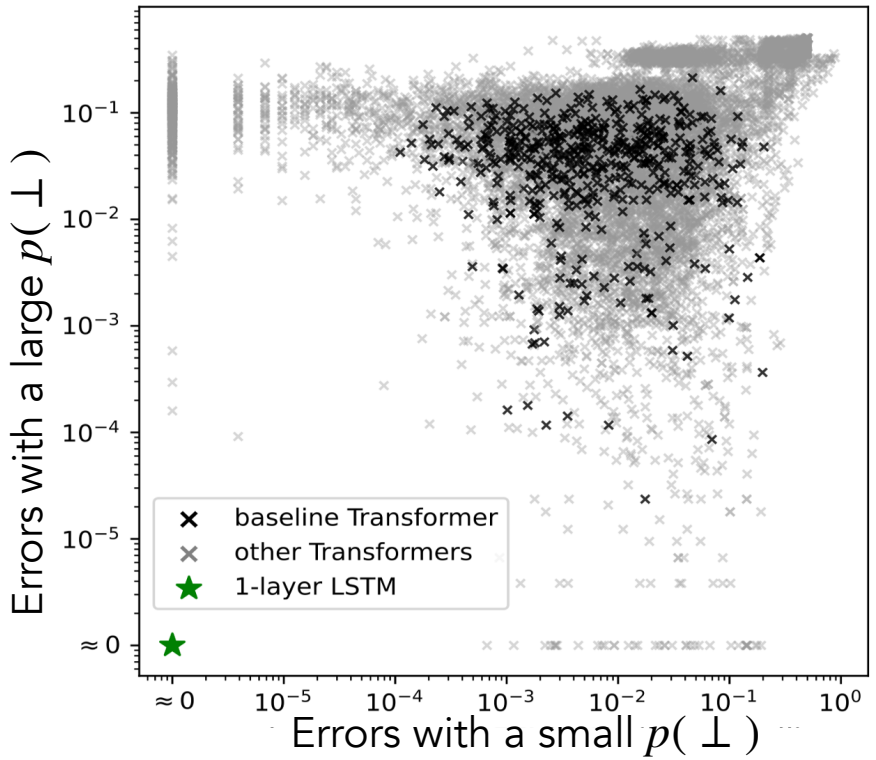

- Irrelevant contexts [Shi et al. 23, Yang et al. 25]

  Jessica is six years older than Claire. In two years, Claire will be 20 years old. *Twenty years ago, the age of Claire's father is 3 times of Jessica's age.* How old is Jessica now?

- Role spoofing attack.

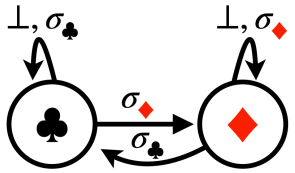



different $p(\perp)$



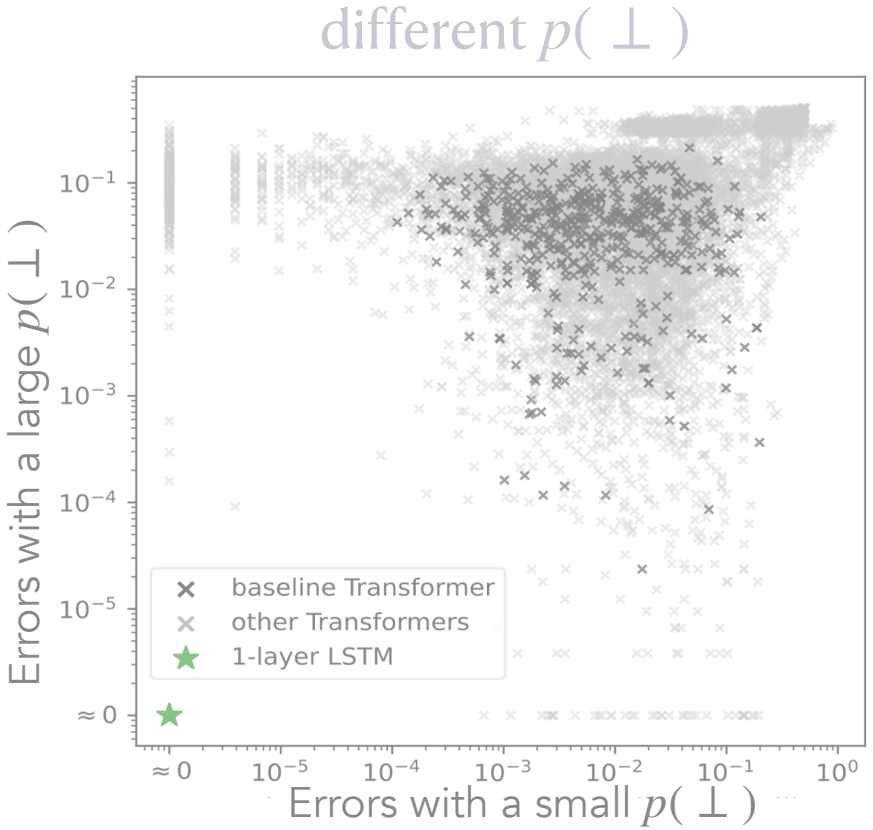
# How well can models *learn* a "base case"?

Memory unit

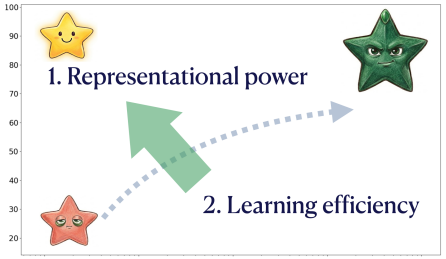



- Errors due to intrinsic architectural limitations *... easy for RNNs.*
  - *Intuition*: "diluted" attention.
- Mitigation: data/architecture changes.

[*ALSK 26 — stay tuned!*]

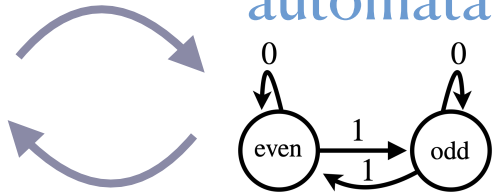


# Lifting the scaling curve




## Part 1: Representational power

Parallel compute for sequential reasoning



automata



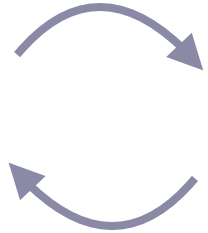

*diagnosing models ← shallow constructions*

LAGKZ: Transformers Learn Shortcuts to Automata. *ICLR 23 (Oral)*.

LAGKZ: Exposing Attention Glitches with Flip-Flop Language Modeling. *NeurIPS 23 (Spotlight)*.

## Part 2: Learning efficiency

Improving distillation



sparse parity

$x = 1\ -1\ -1\ 1\ -1\ 1$

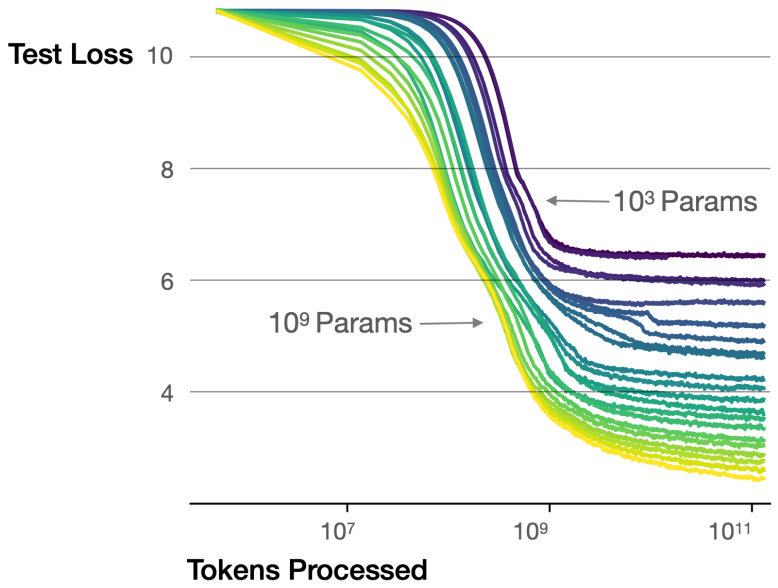
*how to choose ← why teacher choice matters*

P*L*MRG: Progressive distillation induces an implicit curriculum. *ICLR 25 (Oral)*.

PLMKG: In Good GRACEs: Principled Teacher Selection for Knowledge Distillation. *ICLR 26*.

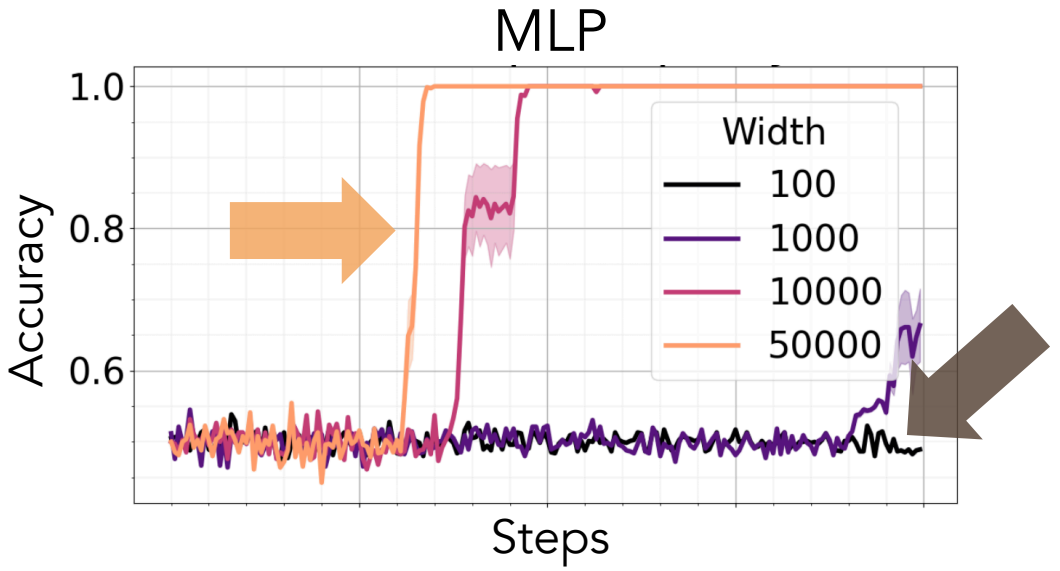
# Smaller models learn more slowly

## Reality (LLM pretraining)




[Kaplan et al. 20]

## Sandbox (sparse parity)




*1) SGD as a statistical query (SQ) algo;*

*2) parity is SQ hard.* [Kearns 98, FGV17, AKMSS21, EGKMZ23]

# Improving smaller models with distillation

Distillation: train small models better, by leveraging powerful pretrained models.

"student" "teacher"

- Student learns from teacher's output:

Supervised learning: $\ell(f(x), y)$ Distillation from $f_T$: $\ell(f(x), f_T(x))$

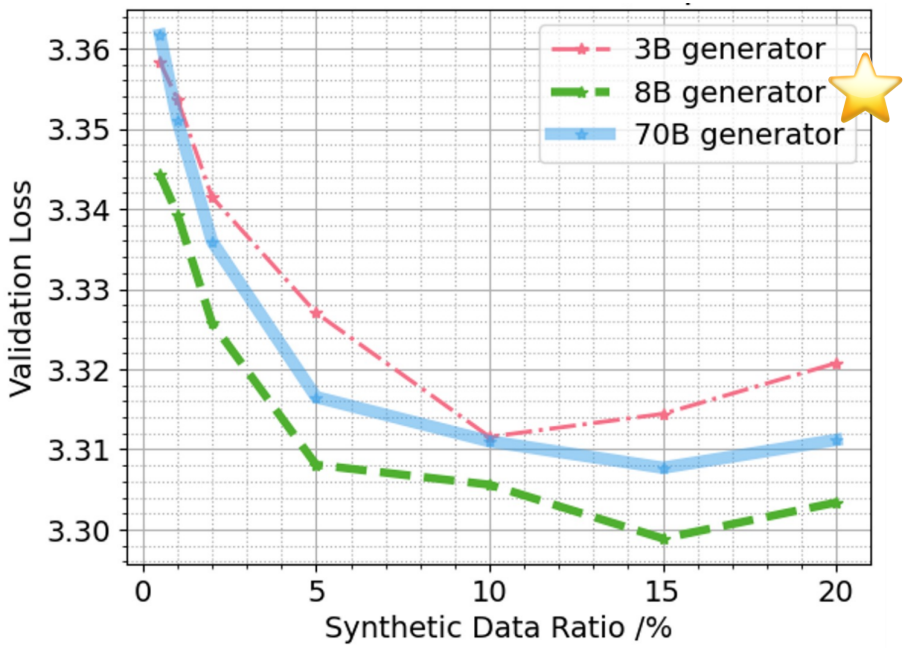
*e.g., logits, generations*

| Model | AIME 2024 | | MATH-500 | GPQA Diamond | LiveCodeBench |
|---|---|---|---|---|---|
| | pass@1 | cons@64 | pass@1 | pass@1 | pass@1 |
| **QwQ-32B-Preview** | 50.0 | 60.0 | 90.6 | 54.5 | 41.9 |
| **DeepSeek-R1-Zero-Qwen-32B** | 47.0 | 60.0 | 91.6 | 55.0 | 40.2 |
| **DeepSeek-R1-Distill-Qwen-32B** | **72.6** | **83.3** | **94.3** | **62.1** | **57.2** |

[DeepSeek R1 report]

*(see also: Gemma, Qwen3, Muse glimmer)*

# Stronger model ≠ better teacher [Mirzadeh et al. 19]




[Kang et al. 25]

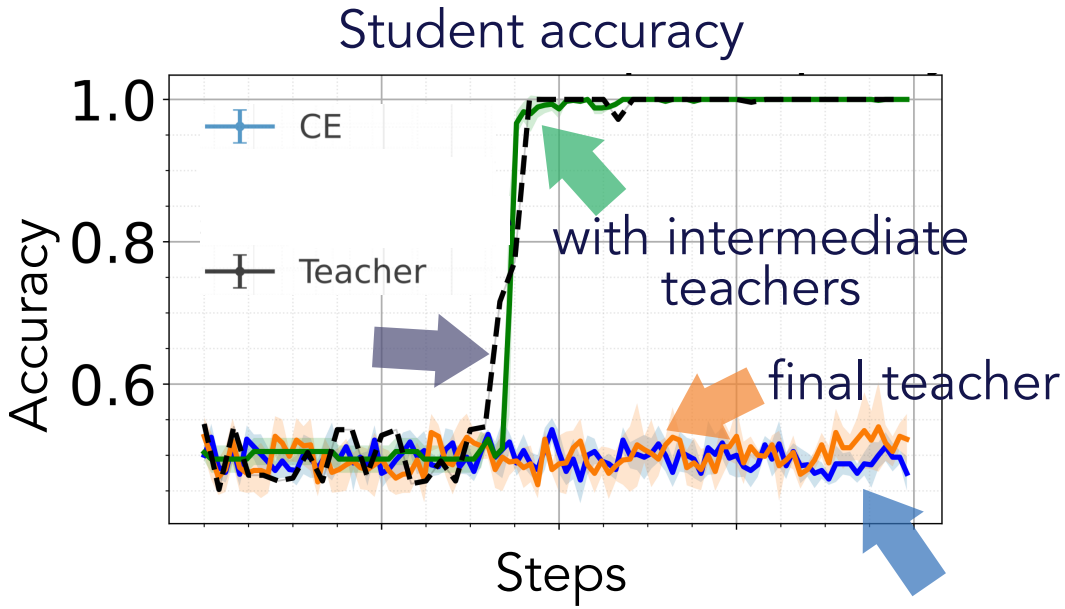
Mitigate with proper **teacher choice**:

1. Distilling with intermediate checkpoints.
   - Provable improvement on sparse parity.
2. Selecting from a pool of open-source models.
   - A lightweight GRACE score for math SFT.

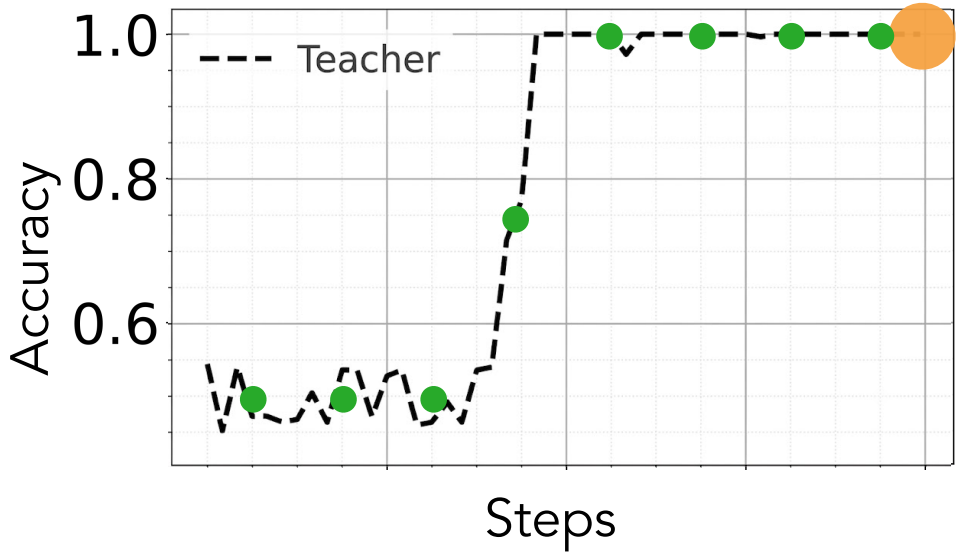
# Distilling with intermediate teachers

$(d, k)$-sparse parity: $x = 1\ {-1}\ {-1}\ 1\ {-1}\ 1\ 1\ 1\ {-1}\ 1 \in \mathbb{R}^d \rightarrow y = \prod_{i \in S} x_i = 1$

$S \ldots |S| = k$



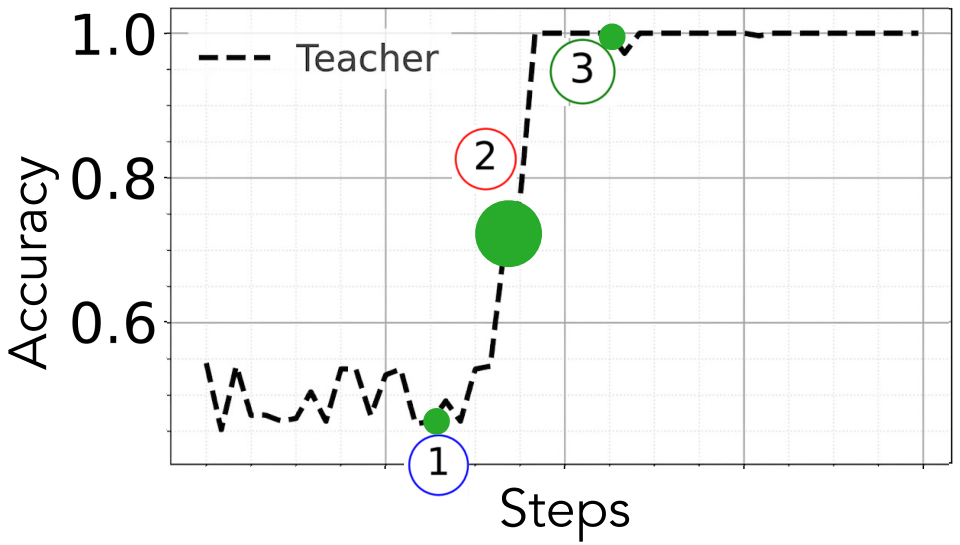

$f_T(x) \approx y$

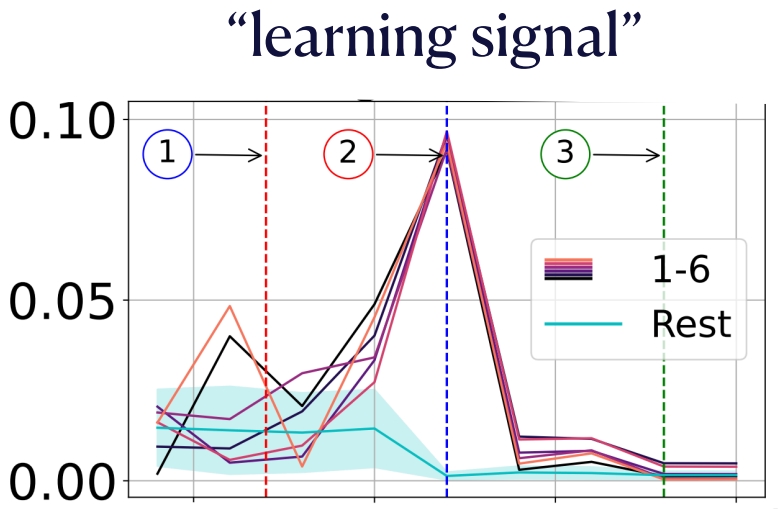


# Signals from intermediate teachers

$$y = \prod_{i \in S} x_i$$

Intermediate checkpoints provide an implicit curriculum. ... ● **reveals** $S$.

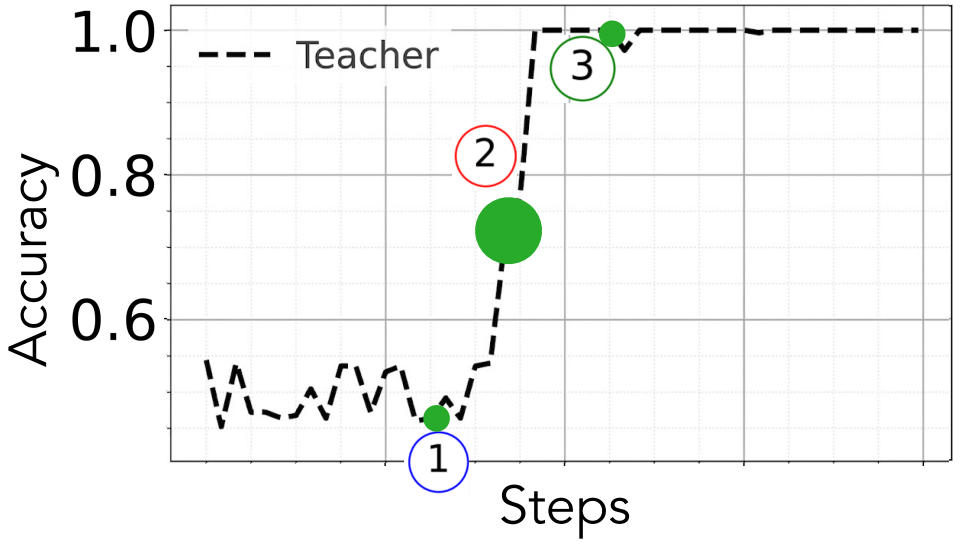


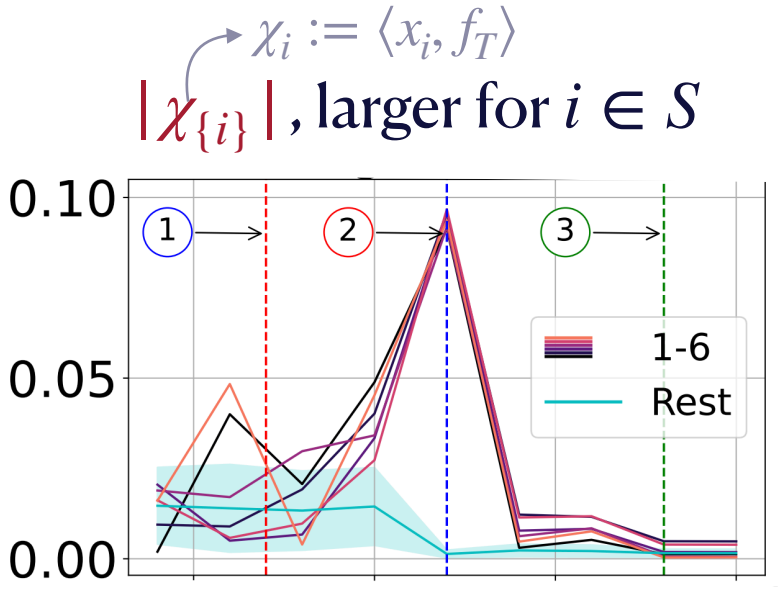



# Signals from intermediate teachers

$$y = \prod_{i \in S} x_i$$

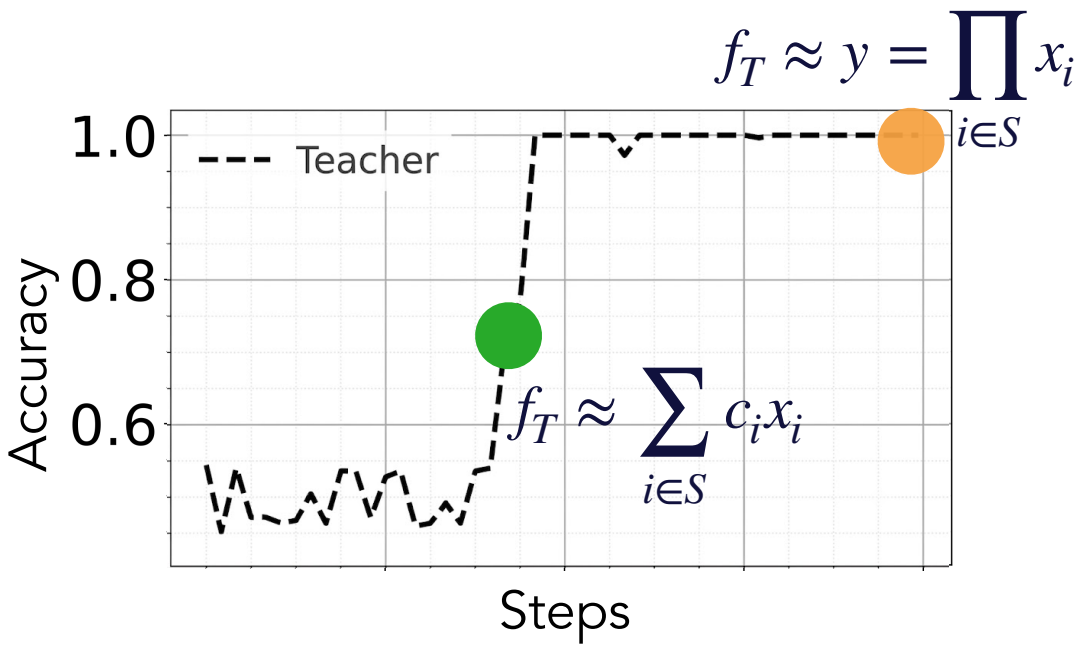
Intermediate checkpoints provide an implicit curriculum. ... ● **reveals** $S$.



# Signals from intermediate teachers

Intermediate checkpoints provide an implicit curriculum. ... ● **reveals** $S$.




● : learnable in $\Omega(d^k)$ steps.

● → ● : learnable in $\tilde{\Theta}_{k,\epsilon}(d^2)$ steps.

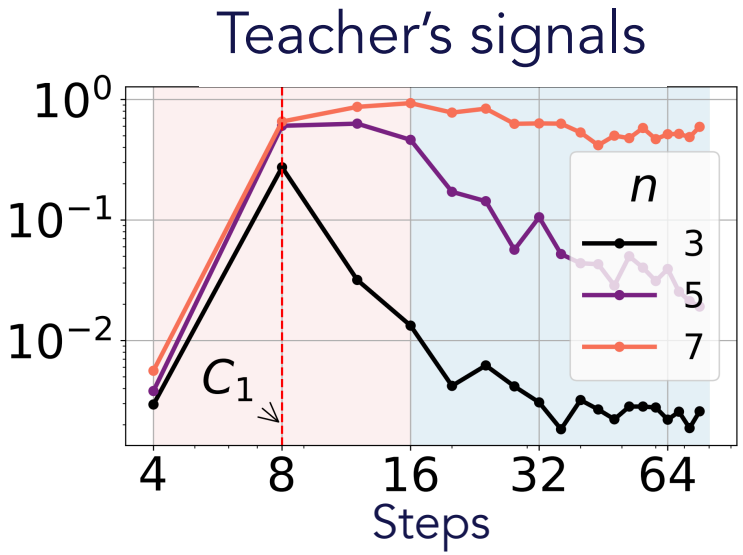
Proof idea: *fewer* steps to learn *lower-degree* polynomials

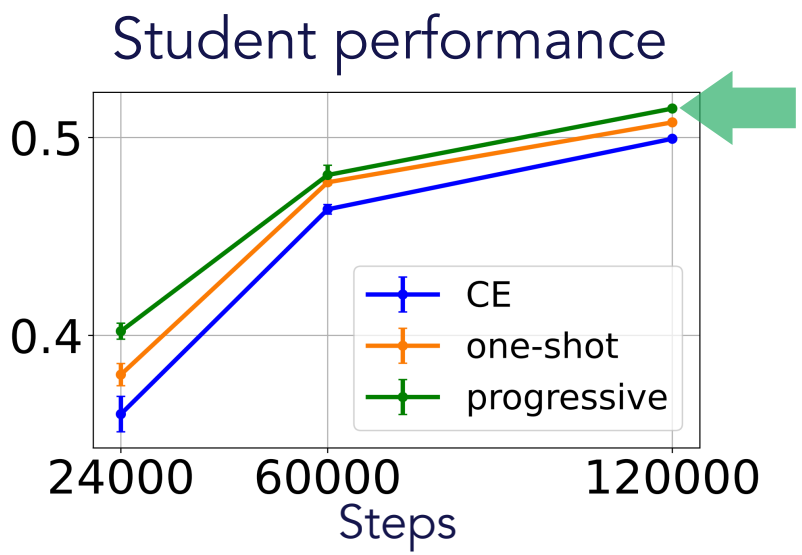
[Edelman et al. 22, Abbe et al. 23]

# Signals from intermediate teachers

Intermediate checkpoints provide an implicit curriculum.

*Same observations for language models*



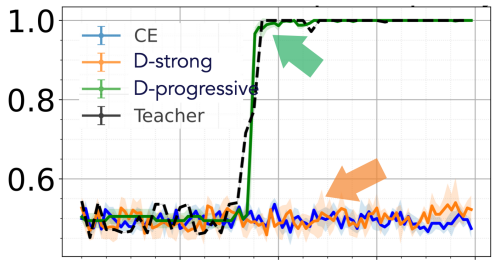

*Related: sensitivity [Vasudeva et al. 24]*

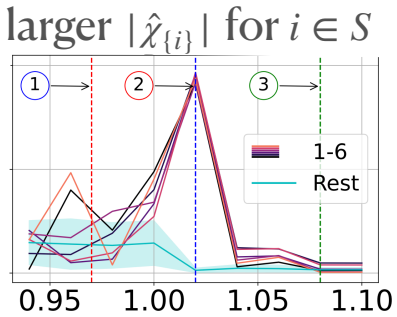
# Stronger model ≠ better teacher [Mirzadeh et al. 19]

Mitigate with proper **teacher choice** → should be **student-aware**.

1. Intermediate checkpoints.
   - As implicit curriculum (sparse parity)
   - Motivate temporal pooling (MUSE Glimmer)

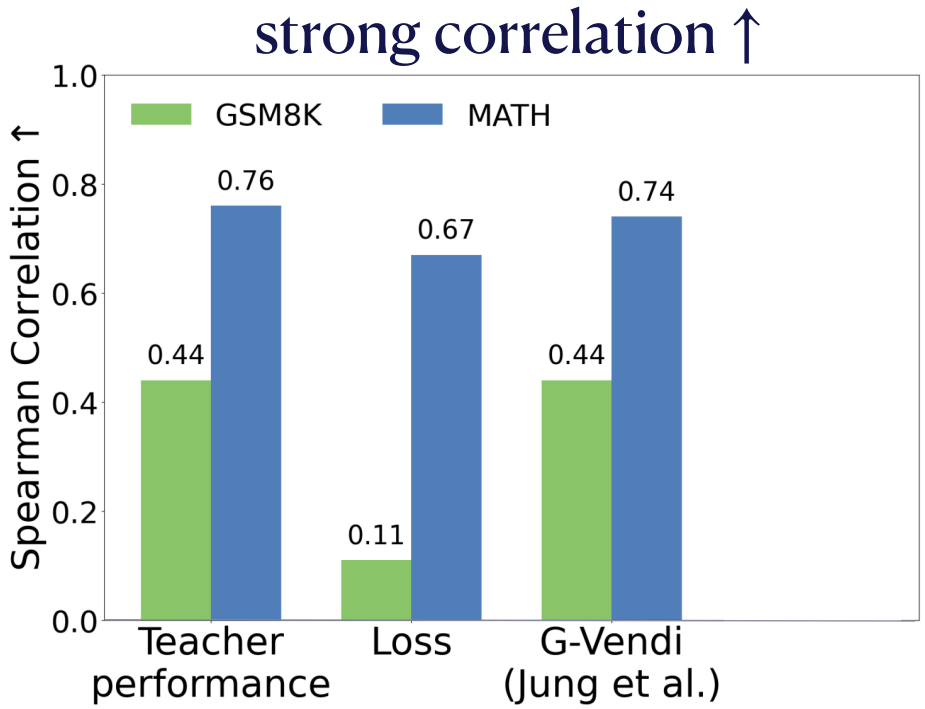
2. From a pool of open-source models.
   - A lightweight GRACE score for SFT.



# LLM teacher selection

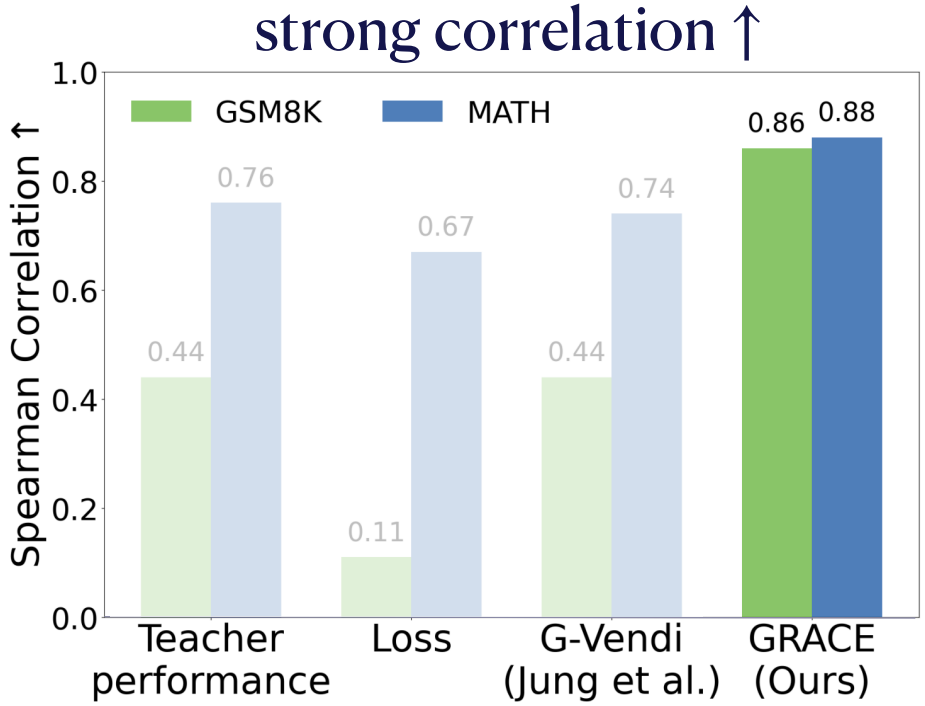
*A student-aware score?*



# GRACE: a score for LLM teacher selection

(GRAdient Cross-validation Evaluation)

- Indicative of the student's performance.
- Informing distillation practices.
  - generation temperature?
  - teacher under a constrained size?
  - teacher within a family?



# Designing GRACE (GRAdient Cross-validation Evaluation)

Commonly used for data selection.

Student's **gradients** on teacher's (unverified) **generations**.

- Computationally light-weight ... parallel forward passes.
- Black-box access suffices ... e.g. API access.
- Applicable across model families ... no tokenizer issues.

**GRACE**: gradient **norm**, weighted by the **spectrum** of normalized gradients.
(G-Vendi)

# GRACE for LLM teacher selection

Task: **math reasoning** (GSM8K, MATH).

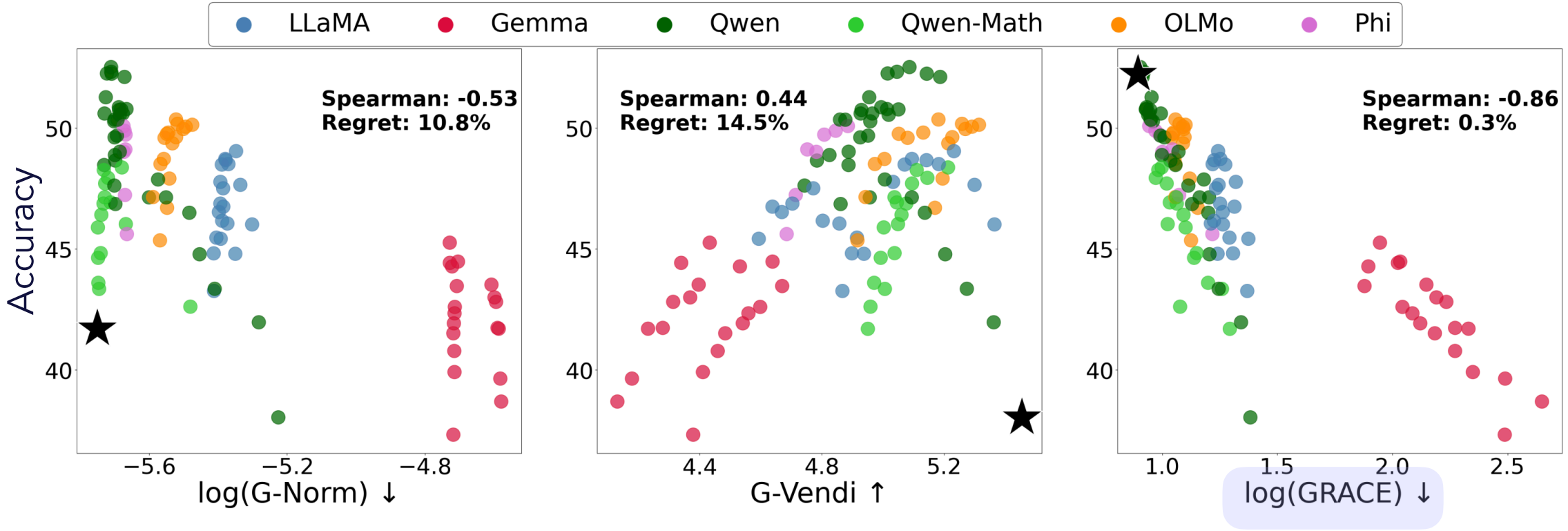
Students: Llama-1B / OLMo-1B / Gemma-2B (GSM); Llama-3B (MATH).

Teachers: **15 models** in Gemma, Llama, OLMo, Phi, Qwen (Math),
at temperatures [0.3,1].

- Scored by GRACE and baselines (norm only, spectrum only).
- Score quality: **correlation** to & **regret** of student's accuracy.

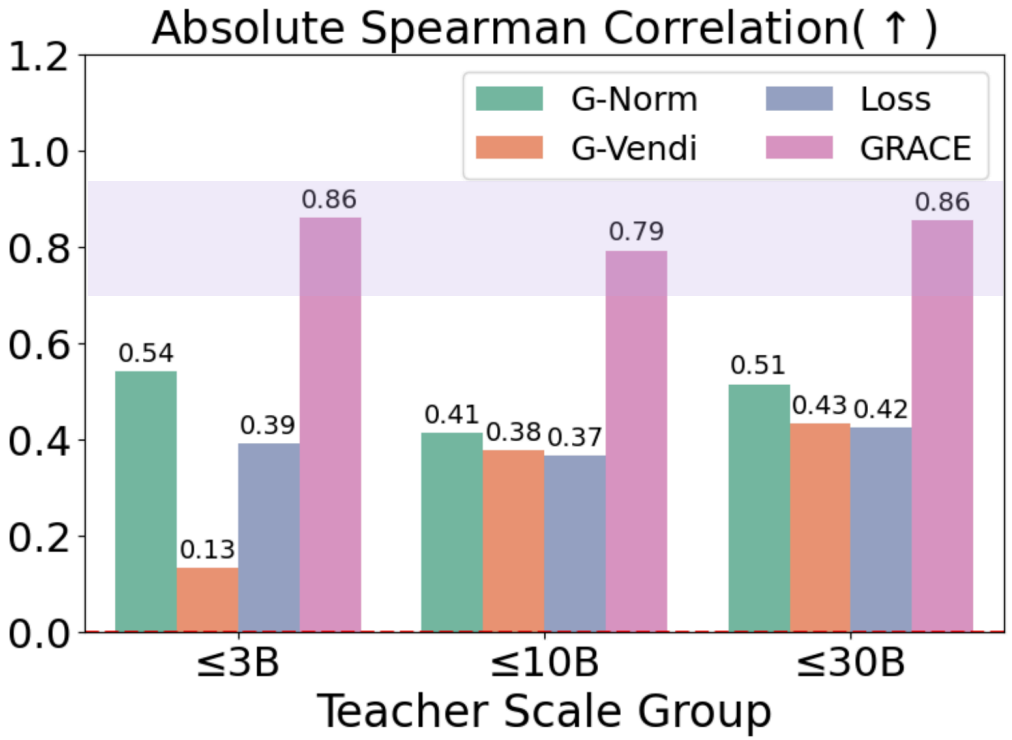
# GRACE: high correlation, low regret

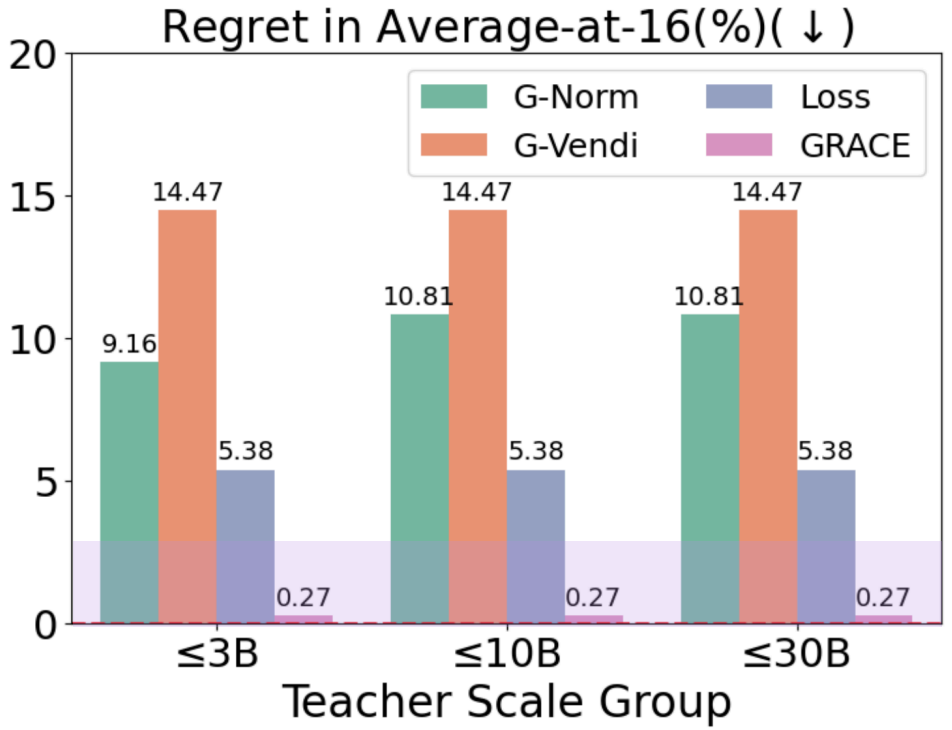
LLaMA-1B on GSM8K *Similar results for other students and datasets (MATH and OOD).



# GRACE guides distillation practice

Example: selecting teacher under size constraints.

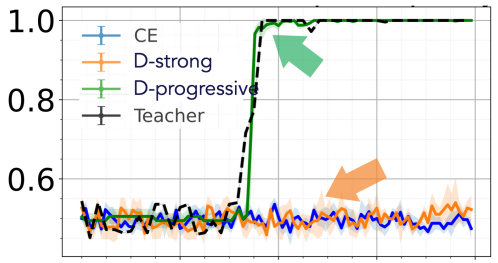


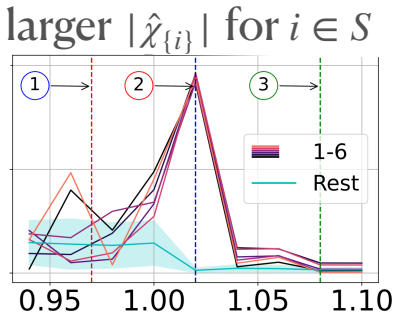



# Stronger model ≠ better teacher [Mirzadeh et al. 19]

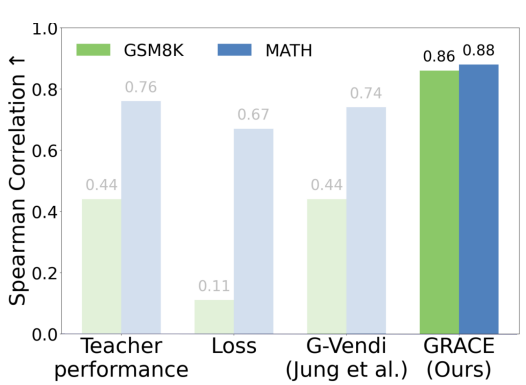
The proper **teacher choice** should be **student-aware**.

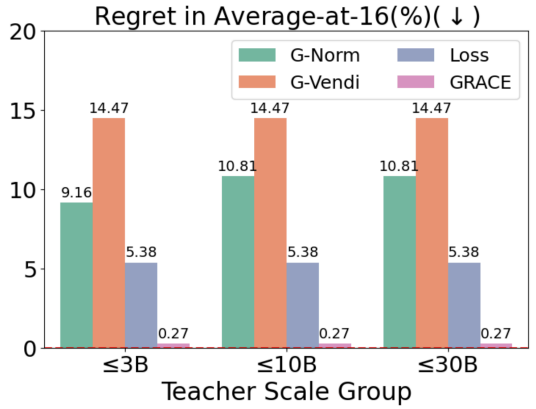
1. Intermediate checkpoints.
   - As implicit curriculum.
   - Motivate temporal pooling.



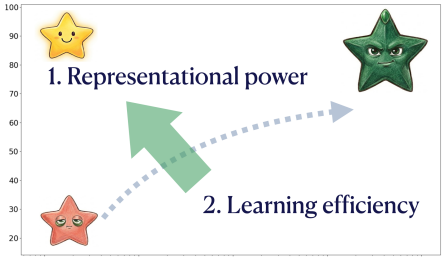




2. From a pool of open-source models.
   - A lightweight GRACE score for math SFT.



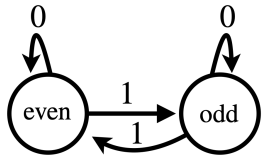
# Lifting the scaling curve




## Part 1: Representational power

Parallel compute for sequential reasoning

automata




Shallow constructions + diagnosis

*o(T)-layer* *OOD failures*

LAGKZ: Transformers Learn Shortcuts to Automata. *ICLR 23 (Oral)*.

LAGKZ: Exposing Attention Glitches with Flip-Flop Language Modeling. *NeurIPS 23 (Spotlight)*.

## Part 2: Learning efficiency

Improving distillation

sparse parity

$x = 1\ -1\ -1\ 1\ -1\ 1$

Student-aware teacher choice.

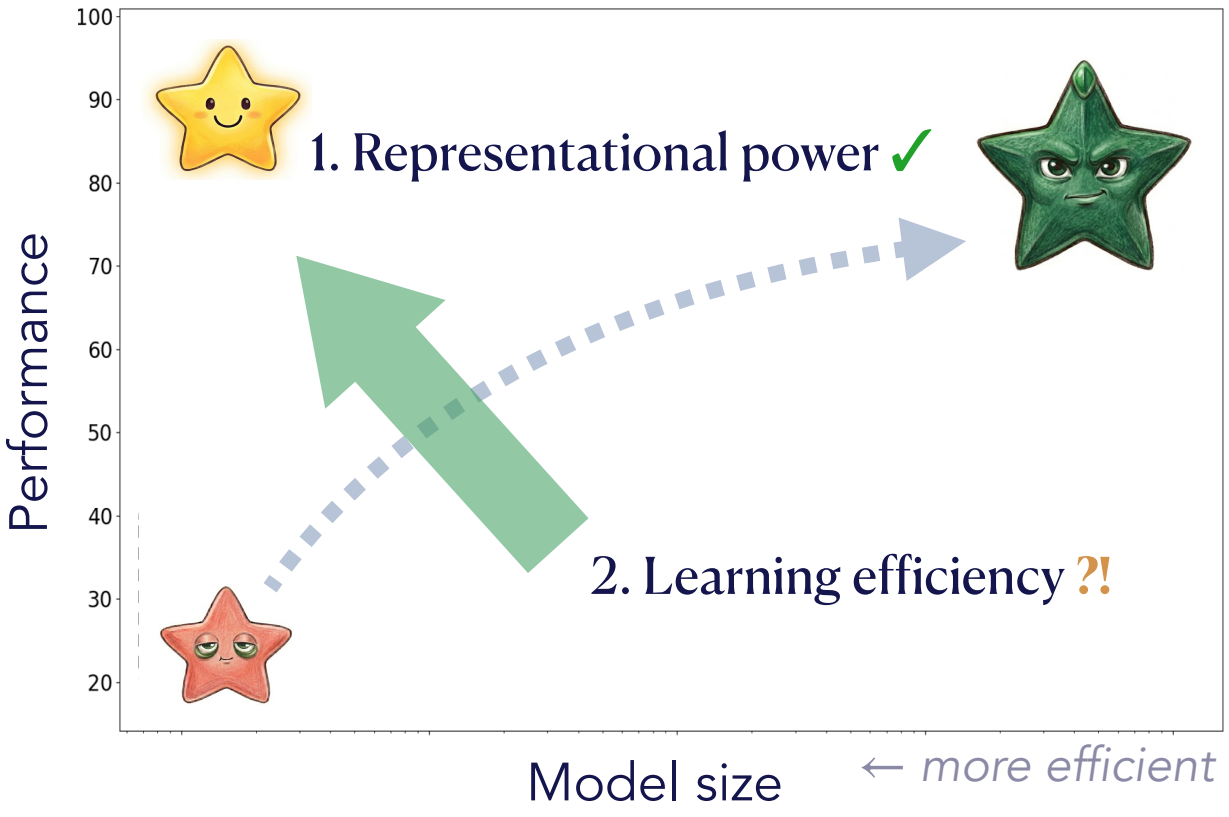
*progressive distill, GRACE*

P*L*MRG: Progressive distillation induces an implicit curriculum. *ICLR 25 (Oral)*.

PLMKG: In Good GRACEs: Principled Teacher Selection for Knowledge Distillation. *ICLR 26*.

# Understanding & Realizing Potentials of Smaller Models

*algorithms matter*




**Expressivity:**

- Architectural alternatives? [ALSK 26]

**Learning:**

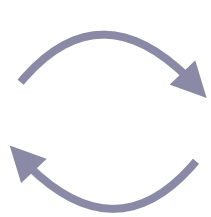
- (On-policy) distillation?
  - Role of non-realizability [ZGKLR 26]
- Data usage?
  - Less data, faster training [LEGL 26]
- Optimizer?
  - Preconditioner design [LBMVAK 25]

**Other axes** ... test-time compute scaling?

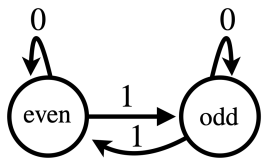
# Understanding & Realizing Potentials of Smaller Models

## Part 1: Representational power

Parallel compute for sequential reasoning



automata

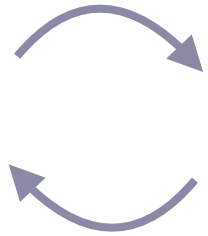



Shallow constructions + diagnosis

*o(T)-layer* *OOD failures*

## Part 2: Learning efficiency

Improving distillation



sparse parity

$x = 1$ -1 -1 1 -1 1

Student-aware teacher choice.

*progressive distill, GRACE*





