Youth in High-Dimensions

ICLR2025

Improving training with progressive distillation

Bingbin Liu
CMU → Simons → **Kempner**

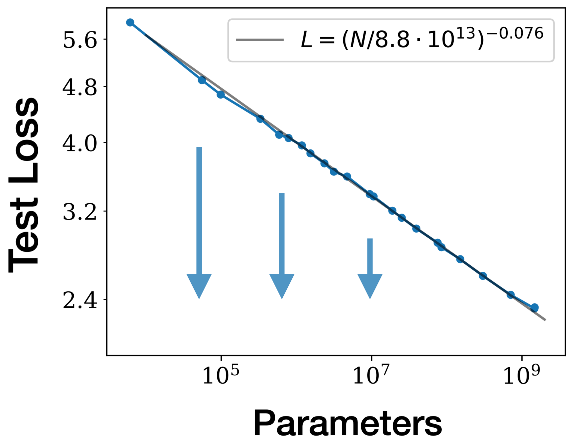
Abhishek Panigrahi

Sadhika Malladi

Andrej Risteski

Surbhi Goel

Better small models?



Challenges mainly in training,
rather than capacity/expressivity.

- Most models are **sufficiently big**.

  e.g. $\Omega(\log T)$ layers [Merrill & Sabharwal 22].

  context length $T = 10^6 \rightarrow$ ~20 layers $\ll$ ~100 layers in practice.

- Other tricks available, e.g. Chain-of-Thought [Li et al. 24].

Better small models?

Train small models better, given big pretrained models?

model compression

e.g. distillation, quantization, pruning

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]

Better small models?

Train small models better, given big pretrained models?

via distillation

Benefit: improved efficiency.

- **Inference**: lower compute cost, while remaining performant.

Better small models?

Train small models better, given big pretrained models?

via distillation

Benefit: improved efficiency.

- **Training**: fewer samples (statistical) / steps (computational).

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

[Hinton et al. 15]

Distillation for better training

ICLR2025

Background: what & how to distill.

- Explanation: *generalization* benefit… limited understanding about ***training***.

Our work: better training via progressive distillation.

- Via an "implicit curriculum."
- Case study (sparse parity) + empirical verification.

Future directions

What is knowledge distillation?

Training a "student" model to match a (trained) "teacher" model.

- Classification, matching outputs (class distributions):

$$L_D(f(x), f_T(x)) = \mathrm{KL}(f_T(x) \| f(x)). \quad f_T(x), f(x) \in \Delta^{C-1}$$

Recall: learning from data:

$$L_{CE}(f(x), y) = -\log[f(x)]_y = \mathrm{KL}(\delta_y \| f(x)).$$

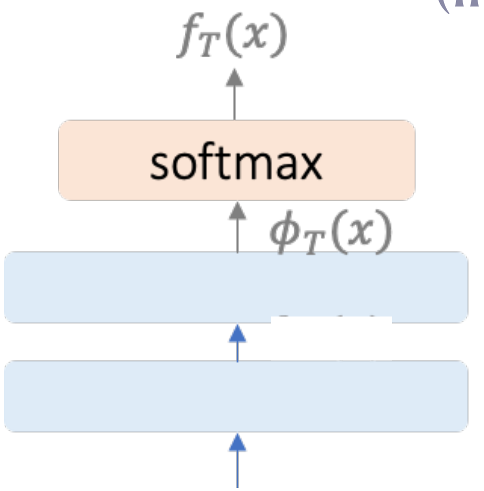
In practice, often use both: $\alpha L_{CE} + (1-\alpha) L_D$.

What is knowledge distillation?

Training a "student" model using a (trained) "teacher" model.

- Classification: matching teacher's output (e.g. class distributions).
- Distribution given by the softmax: $[f(x)]_i \propto \exp(\ \tau^{-1} \cdot [\phi(x)]_i\ )$.

(inverse) temperature



What is knowledge distillation?

Training a "student" model using a (trained) "teacher" model.

- Classification: matching teacher's output (e.g. class distributions).
- Distribution given by the softmax: $[f(x)]_i \propto \exp(\ \tau^{-1} \cdot [\phi(x)]_i\ )$.

** Distillation is fairly general:*

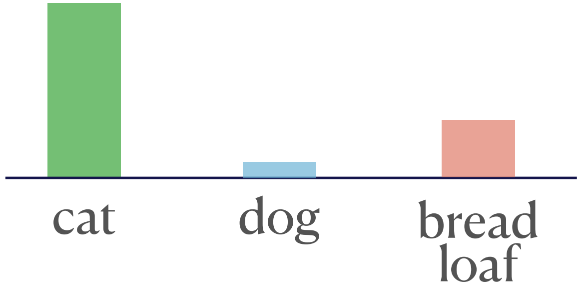
- Big/strong teacher → small/weak student. (today's focus)
- Small/weak teacher → big/strong student (e.g. weak-to-strong)
- Self-distillation (same-sized), many-to-one, ...

Why is distillation helpful?

loss $= \mathrm{KL}(f_T(x) \| f(x))$

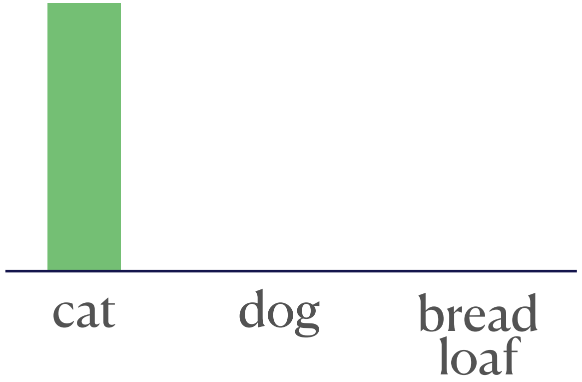
Intuitively: "**richer information**" ... full distribution vs a sample.

- An ideal teacher: $f_T(x) = p^\star(y \mid x)$.



teacher

$p(\cdot \mid x)$



data label

$y \sim p(\cdot \mid x)$

Why is distillation helpful?

$\text{loss} = \text{KL}(f_T(x) \| f(x))$

Intuitively: "richer information" … full distribution vs a sample.

Better generalization: $p^\star(\cdot \mid x)$ leads to a tighter bound [Menon et al. 20].

- Imperfect teacher: bias-variance tradeoff.

Why is distillation helpful?

$\text{loss} = \text{KL}(f_T(x) \| f(x))$

Intuitively: "richer information" ... full distribution vs a sample.

Better generalization: $p^\star(\cdot \mid x)$ leads to a tighter bound [Menon et al. 20].

- Imperfect teacher: bias-variance tradeoff.

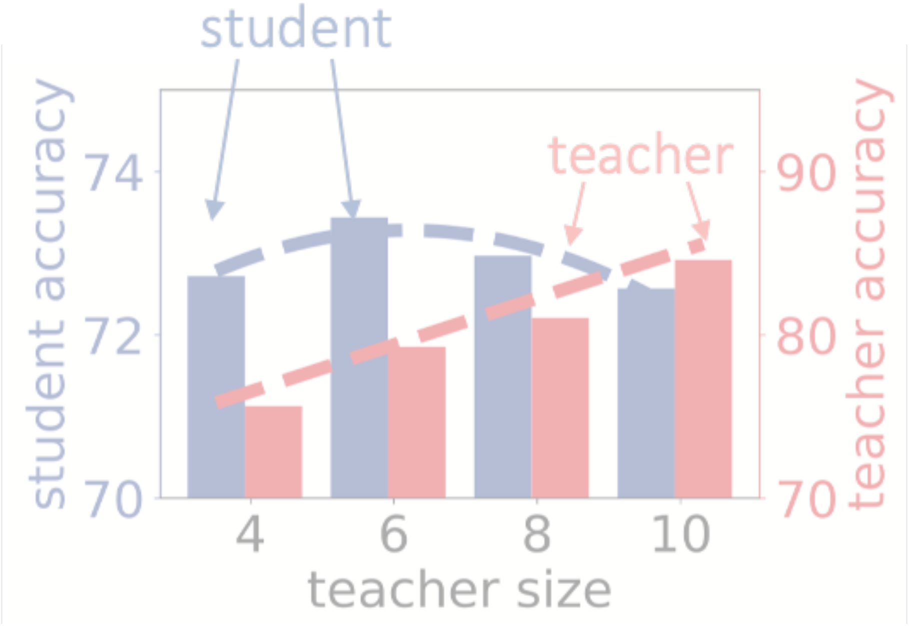
Not the full story — cannot explain:

- Benefit when $p^\star$ is a delta mass? ... i.e. labels = ideal teacher; e.g. sparse parity.
- Better (closer to $p^\star$) teacher $\nrightarrow$ better student?

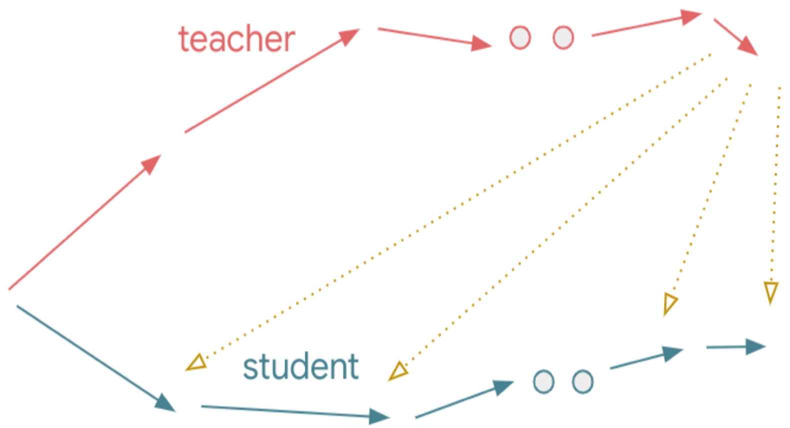
Better teacher $\nrightarrow$ better student

“capacity gap”

(when the teacher is too big / performant)



[Mirzadeh et al. 19]



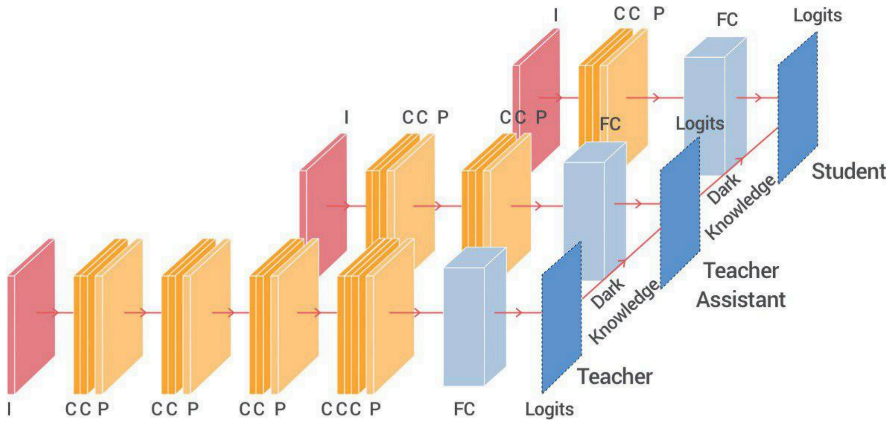
[Harutyunyan et al. 23]

Better teacher ↛ better student

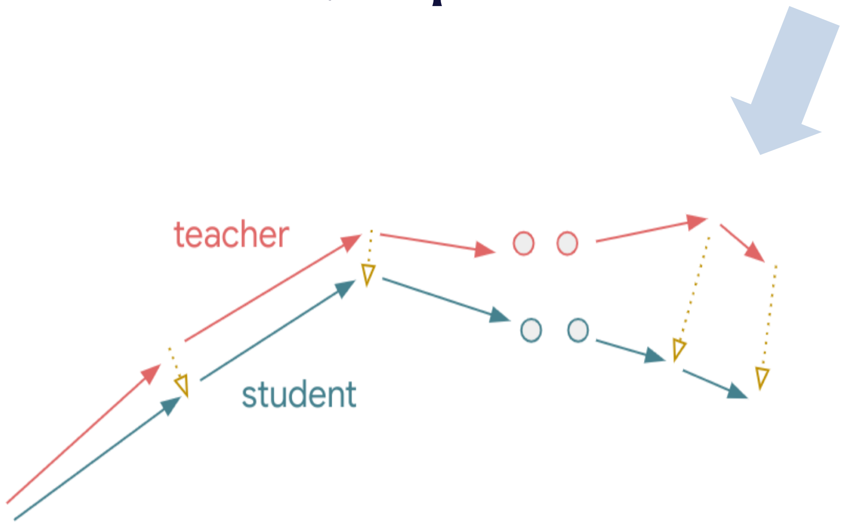
“capacity gap”

(when the teacher is too big / performant)

Bridging the gap with intermediate sizes/steps.



[Mirzadeh et al. 19]

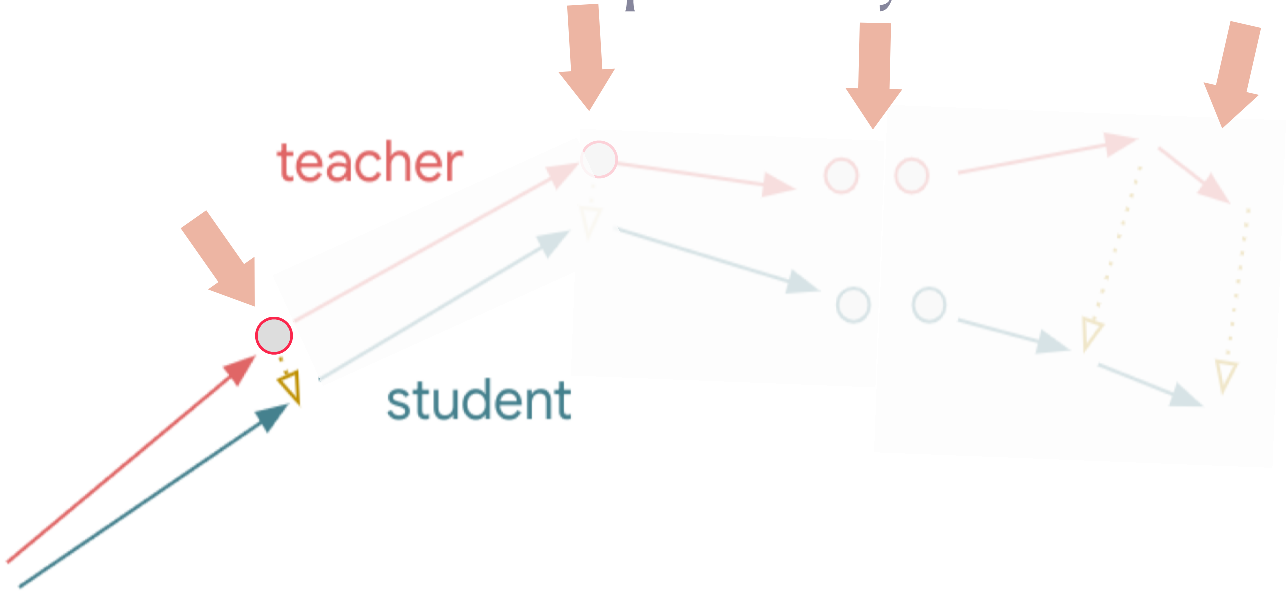


[Harutyunyan et al. 23]

Progressive distillation

Def: student distills sequentially from multiple teacher checkpoints.

- (1-shot) distillation uses the final checkpoint only.



Used in practice: e.g. Gemini-1.5 Flash (from Gemini-1.5 Pro) [Reid et al. 24].

Benefit of progressive distillation?

[Harutyunyan et al. 23]: smaller gap → better generalization (upper) bounds.

Our work: progressive distillation for **faster training**.

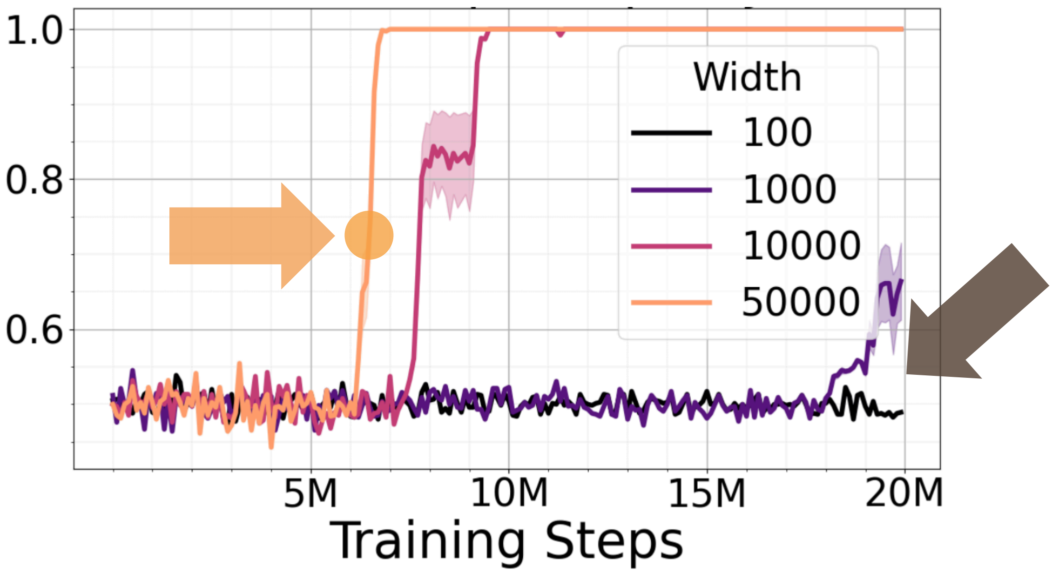
- Case study: **sparse parity** ... prior theory fails to explain the gain.
- Theoretical explanation: reduced sample complexity.
- Empirical validation & more realistic settings (formal and natural languages).

Case study: sparse parity

$$x = 1\ \underbrace{-1\ -1\ 1}_{S}\ -1\ 1\ 1\ 1 \in \mathbb{R}^d,\ |S| = k \ \rightarrow\ y = \prod_{i \in S} x_i = 1$$

- Bigger model trains faster. [Edelman et al. 23]
  - SQ lower bound d^k [Kearns 98]

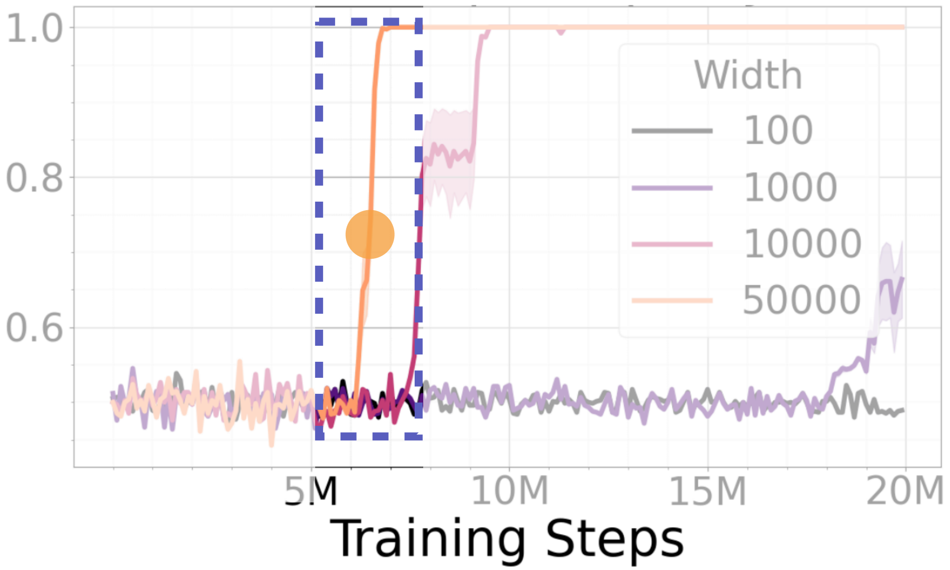
- Our work: Smaller models train as fast, when using intermediate checkpoints.



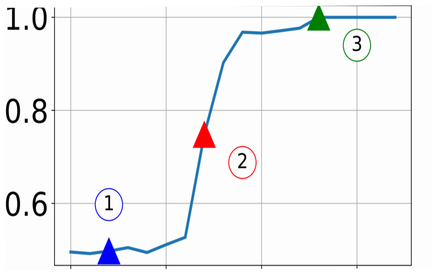
Case study: sparse parity

$$x = 1\ \underbrace{-1\ -1\ 1}_{S}\ -1\ 1\ 1\ 1 \in \mathbb{R}^d,\ |S| = k \ \rightarrow\ y = \prod_{i \in S} x_i = 1$$

- Bigger model trains faster. [Edelman et al. 23]
  - SQ lower bound d^k [Kearns 98]
- Our work: Smaller models train as fast, when using intermediate checkpoints.



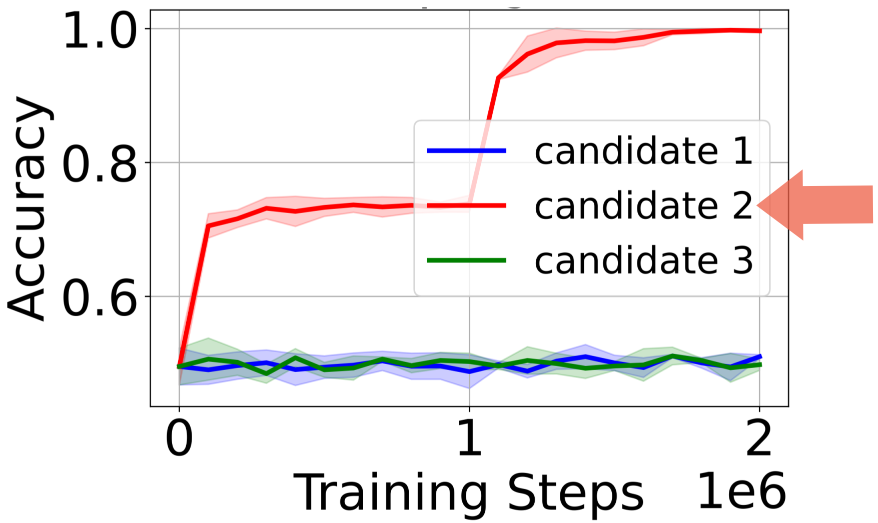
Why intermediate teacher matters?



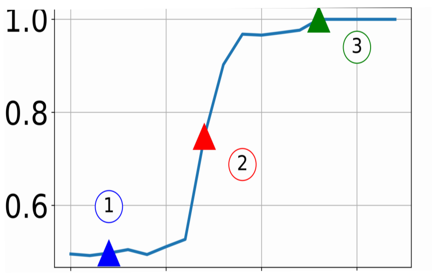
Setup: 2-shot: using 1 intermediate teacher + the final teacher.

Compare 3 teacher checkpoints: before / during / after the phase transition.

① ② ③



Why intermediate teacher matters?

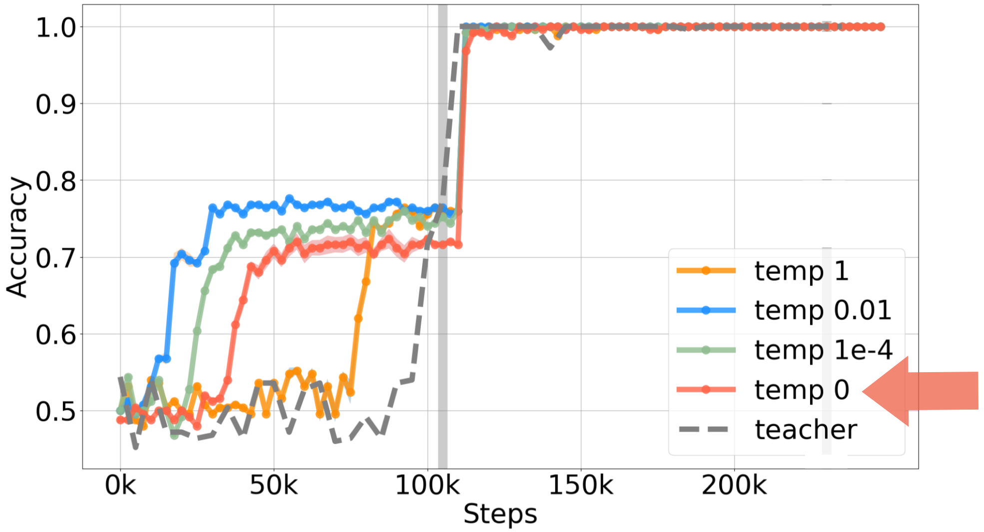


Setup: 2-shot: using 1 intermediate teacher + the final teacher.

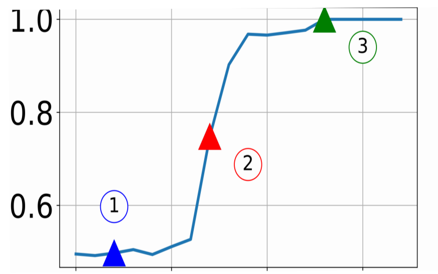
Compare 3 teacher checkpoints: before / during / after the phase transition.

Not due to *"full distribution/soft labels"*: even one-hot supervision is helpful.

- Achieved with a smaller τ.
- Recall: $[f(x)]_i \propto \exp(\tau^{-1} \cdot [\phi(x)]_i)$.



Why intermediate teacher matters?

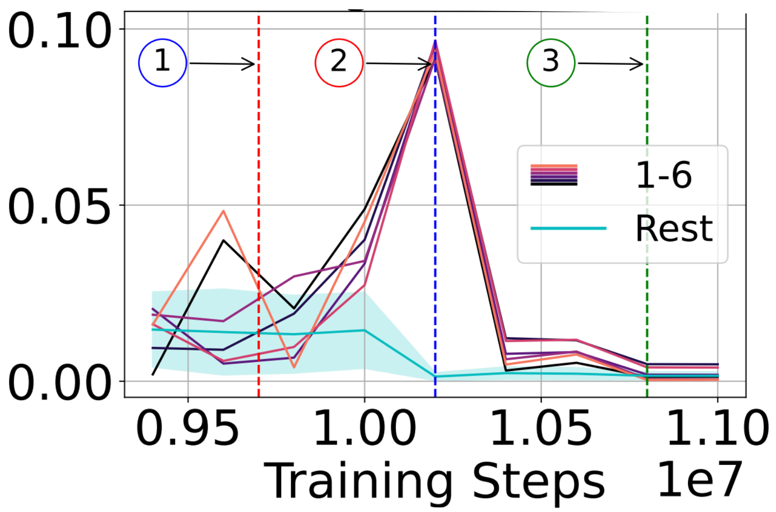


Setup: 2-shot: using 1 intermediate teacher + the final teacher.

Compare 3 teacher checkpoints: before / during / after the phase transition.

True reason: "*extra training signals*" — Implicit curriculum

(certain Fourier coefficients)



Implicit curriculum accelerates training

Case study with **sparse parity**: speedup from "extra training signals."

- What are the signals? ... Fourier coefficients.
- Why are they helpful? ... Lower degree → reduced sample complexity.
- How do they emerge in the teacher? ... Initial population gradient.

(Empirical validation)

Setup

Target: (d, k)-sparse parity: $y = \prod_{i \in S} x_i$, $x \in \{\pm 1\}^d$, $|S| = k$.

Model: 2-layer MLP: $f(x) = \sum_{j \in [m]} a_j \cdot \mathrm{ReLU}(\langle w_j, x \rangle + b_j)$.

Correlation loss $\ell(f(x), y) = -f(x) \cdot y$ or $f_T(x)$ for the student.

- Teacher: 2-phase: 1) one step with a large batch; 2) online SGD.
- Student: 2-shot distillation, from the end of each phase.

What signals: certain Fourier coeffs

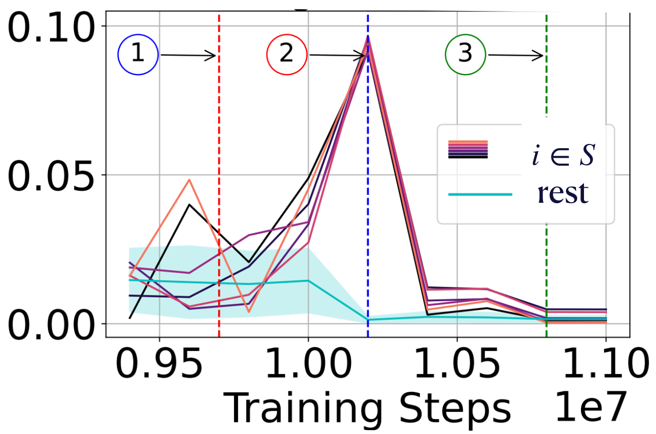
Recall: **Fourier coefficients:** $\hat{f}_{\tilde{S}}(f) = \langle \chi_{\tilde{S}}, f \rangle = \mathbb{E}_x[\chi_{\tilde{S}}(x) \cdot f(x)]$.

- (Fourier basis) $\chi_{\tilde{S}}(x) := \prod_{i \in \tilde{S}} x_i$, for $\tilde{S} \subset [d]$. ... natural for parity: $y = \chi_S$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- Checkpoint 2: $f_T(x) \approx \sum_{i \in S} c_i x_i$

helpful "extra signal"



Why implicit curriculum accelerates training

Fewer samples to learn **lower-degree** monomials [Edelman et al. 22, Abbe et al. 23].

- Learning from $y = \chi_S(x) \rightarrow \Omega(d^{k-1})$ samples.
- Learning from $\sum_{i \in S} c_i \chi_{\{i\}} \rightarrow O(d^2)$ samples.

  $\tilde{O}_{k,\epsilon}(d^2)$ for 2-shot distillation.

2-shot distillation: 1) learn S from $\sum_{i \in S} c_i \chi_{\{i\}}$; 2) compute χ_S given S.

How implicit curriculum arises

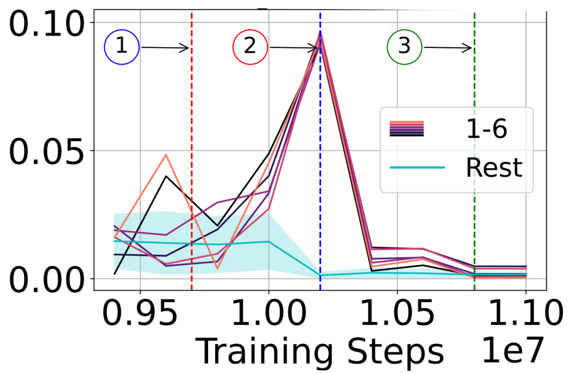
Initial **population gradient** reveals S [Edelman et al. 22].

- Consider a single neuron $w \in \mathbb{R}^d$, its gradient coordinates satisfy
  - Intuition: $|g_j|$ depends on $\hat{f}_{S\setminus\{i\}}$ or $\hat{f}_{S\cup\{j\}}$.

(γ_k : Fourier gap)

$$|g_i| \geq |g_j| + \gamma_k,\ i \in S, j \notin S\,.$$

In support → large gradients



Implicit curriculum accelerates learning

Case study with **sparse parity**: speedup from "extra training signals."

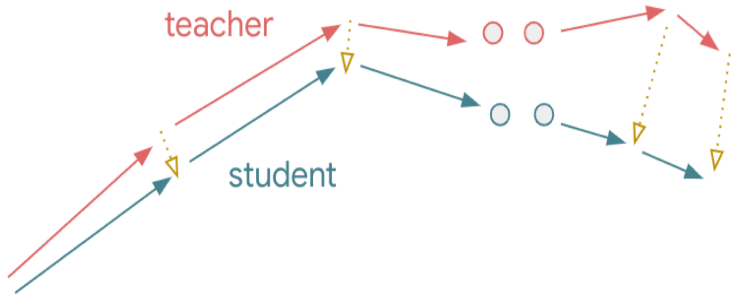
- **What** the curriculum is: *deg-1 monomials*, i.e. $x_i, i \in S$.
- **Why** it is helpful: sample complexity $\Omega(d^{k-1}) \to \tilde{\Theta}_{k,\epsilon}(d^2)$.
- **How** it emerges: initial *population gradient* reveals the support.

Implicit curriculum: a helpful decomposition.

Experiments

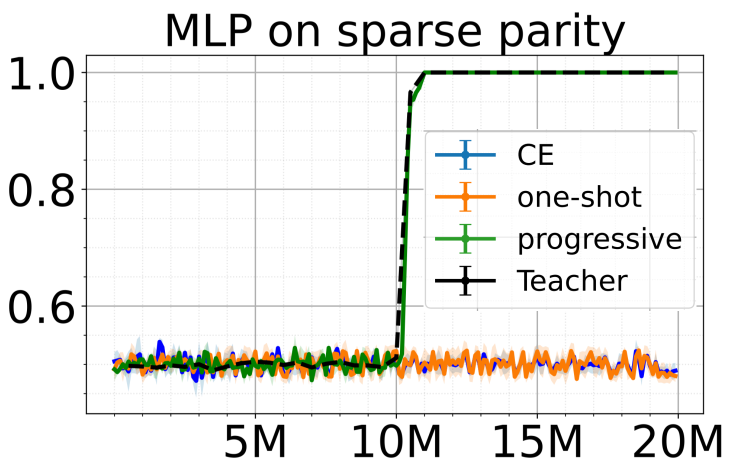
Implicit curriculum for parity and formal/natural languages

Progressive distillation

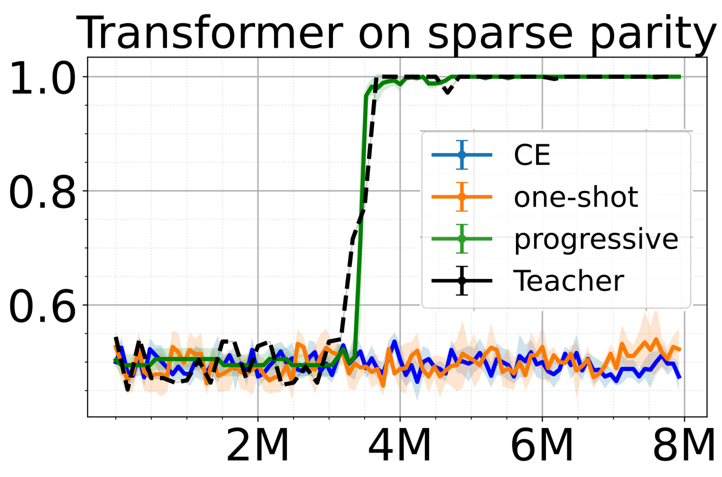


Teacher at fixed intervals (rather than 2-shot).

[Harutyunyan et al. 23]



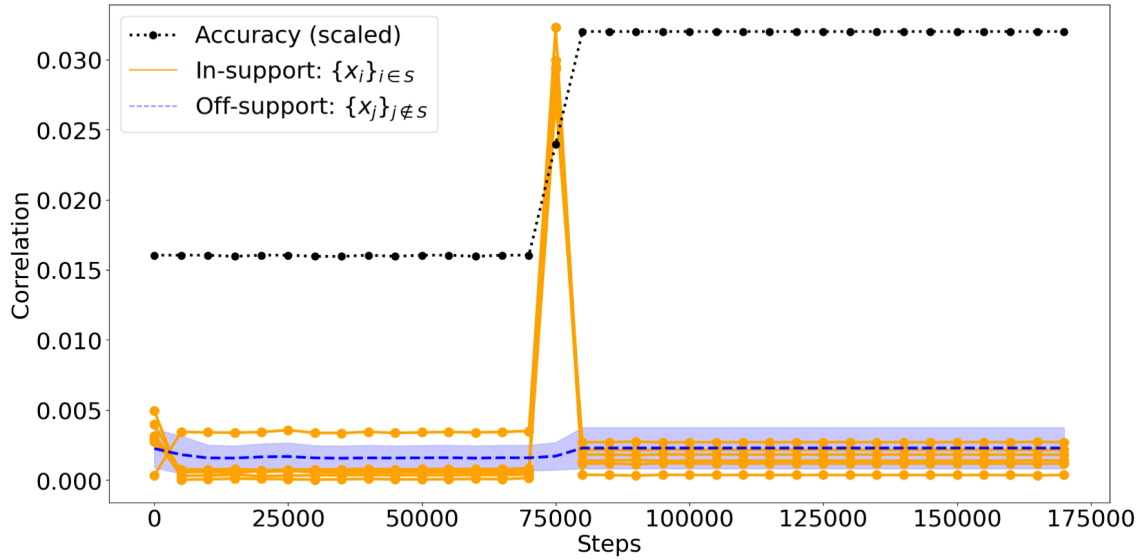
at 100k intervals



at 10k intervals

Transformer on sparse parity

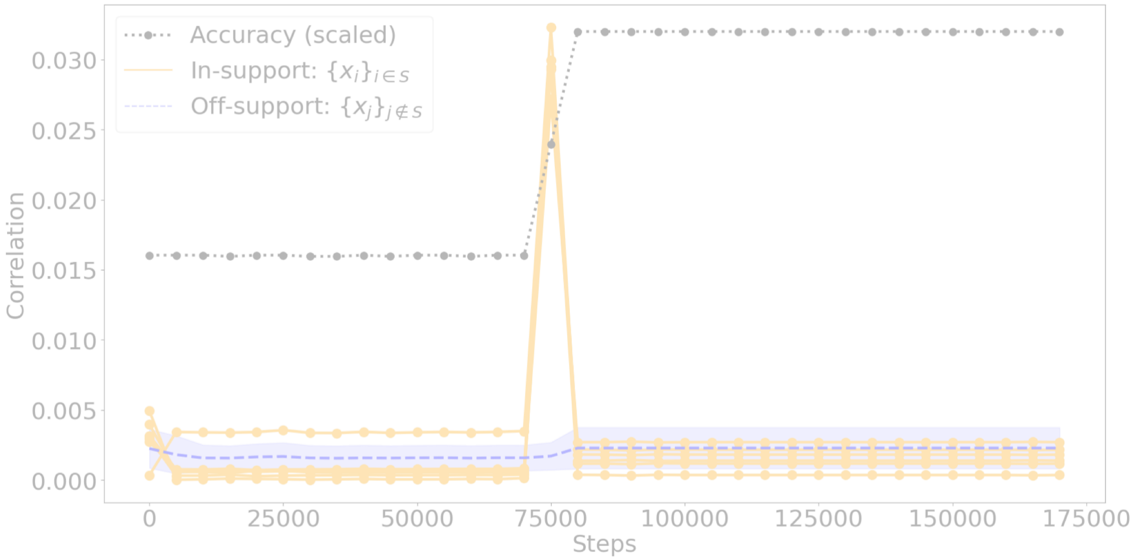
1. **Implicit curriculum** emerges: Higher $\hat{f}_{\{i\}}$ for $i \in S$.

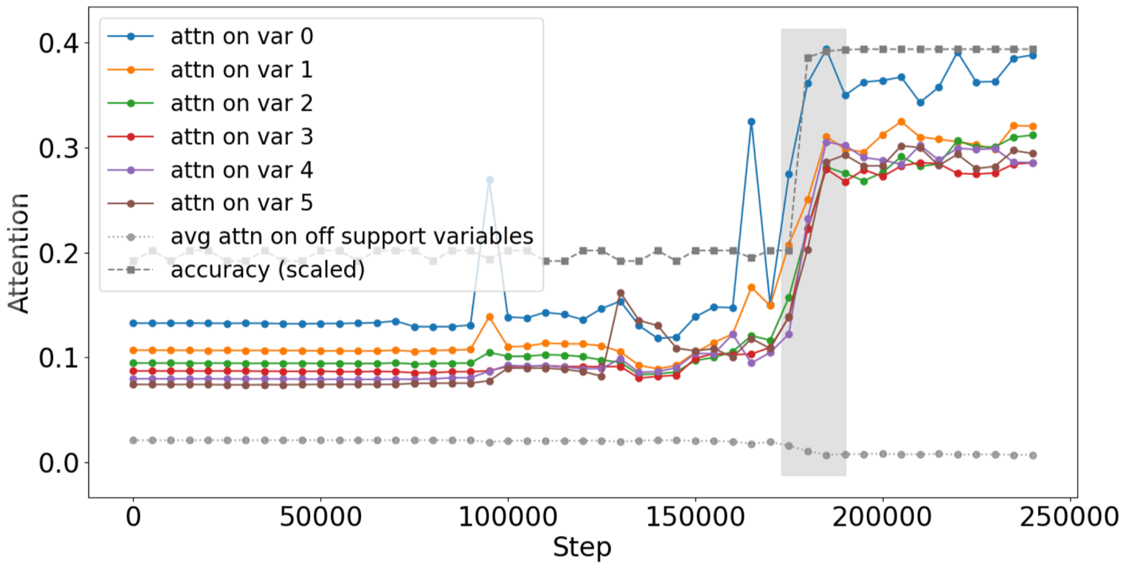


Transformer on sparse parity

1. Implicit curriculum emerges: Higher $\hat{f}_{\{i\}}$ for $i \in S$.

2. **More attention weight** on S.

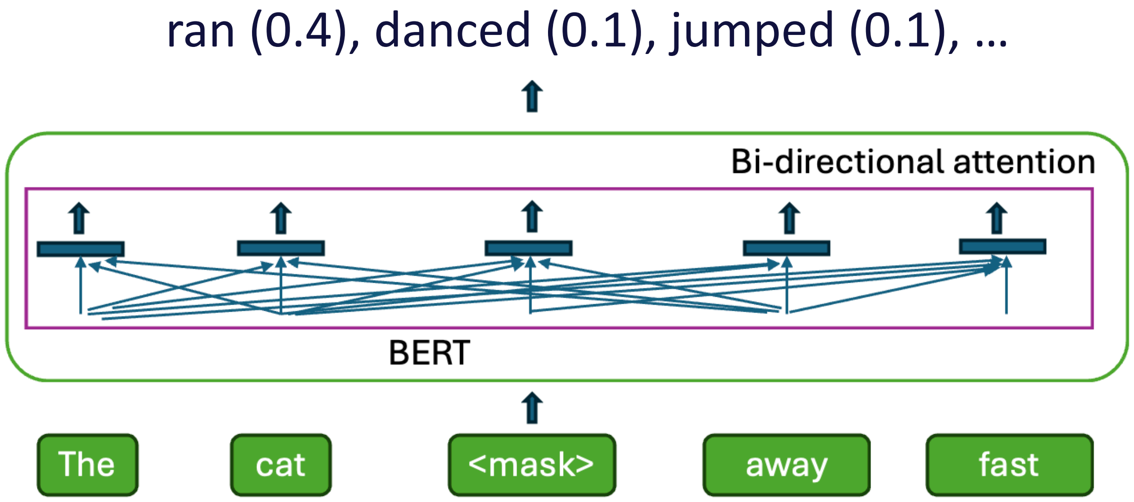




Beyond sparse parity — formal languages

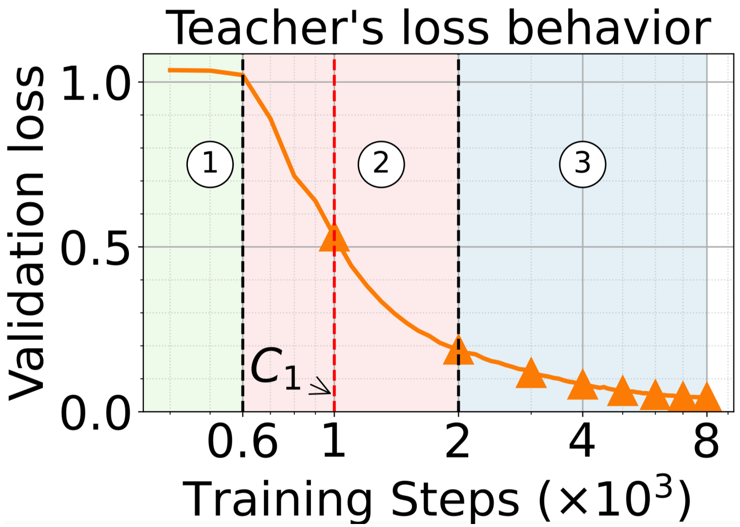
Masked prediction on **PCFG.**

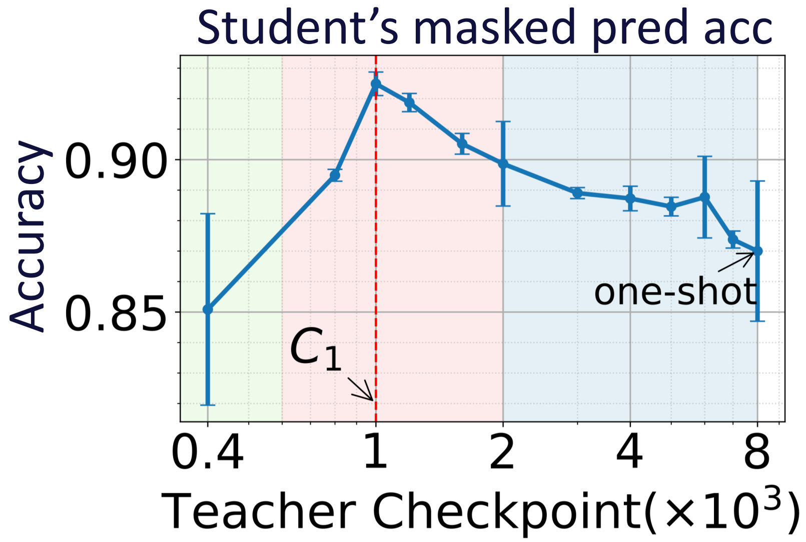
(probabilistic context-free grammar, e.g. [Allen-Zhu & Li 23])



Beyond sparse parity — formal languages

Masked prediction on PCFG: **intermediate checkpoint** helps.



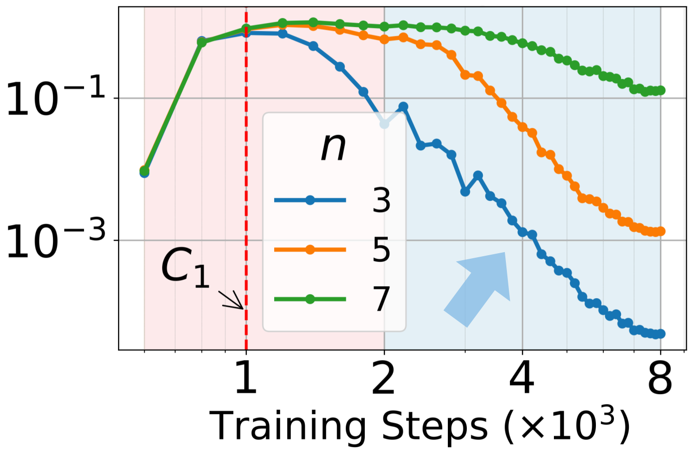


Beyond sparse parity — formal languages

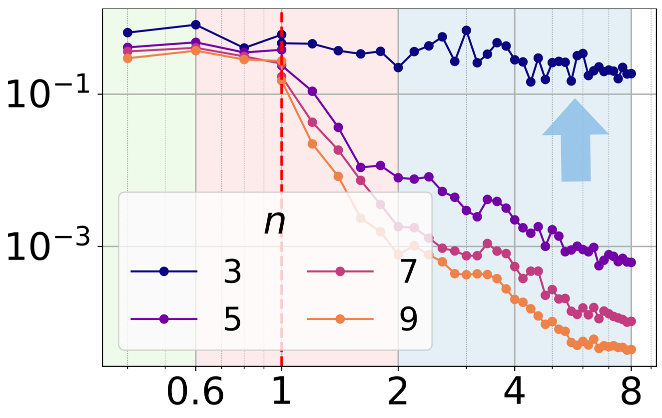
Masked prediction on PCFG: an **implicit curriculum** exists.

n-gram curriculum with an increasing *n*.

“sensitivity” to *n*-gram



“distance” to *n* -gram

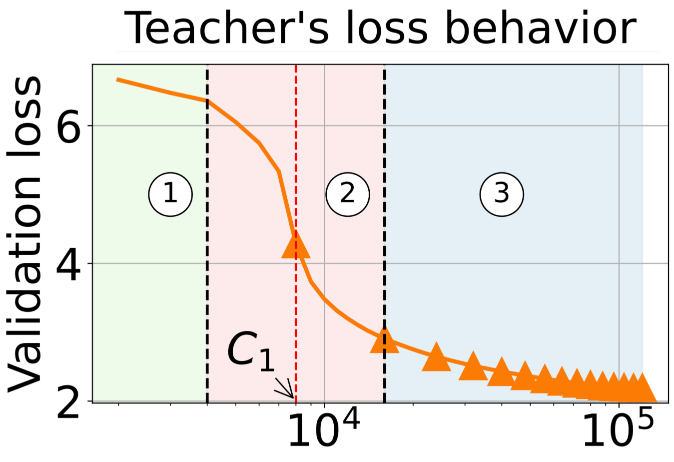


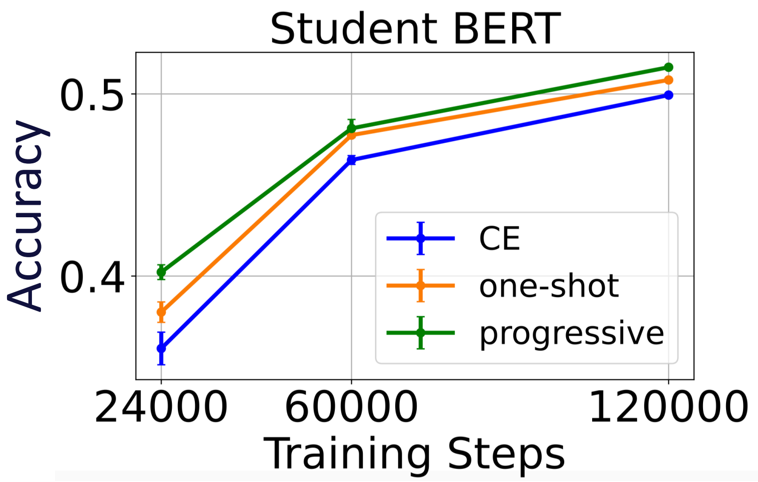
Larger *n* (higher sensitivity/more global) is harder to learn [Abbe et al. 23,24; Vasudeva et al. 24].

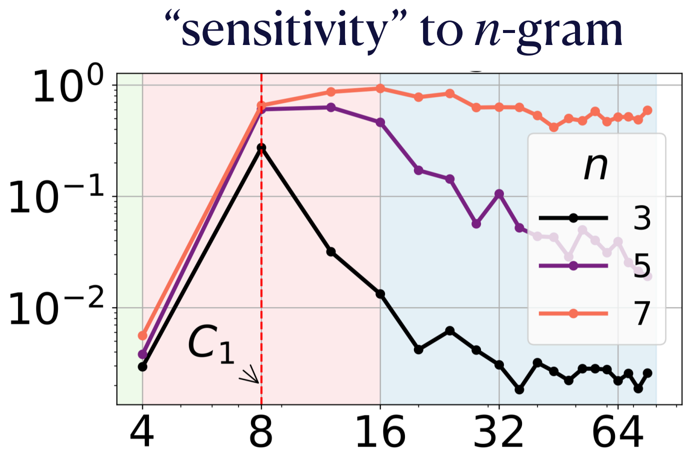
Beyond sparse parity — natural languages

n-gram curriculum for Wikipedia and Books:

- Similar results for masked prediction and next-token prediction.







Progressive distillation accelerates training

(Prior work: *generalization* benefits from full distribution/soft logits)

Intermediate checkpoints provide an implicit curriculum.

- Explains why better teacher $\nrightarrow$ better student ("capacity gap").
- Case study on sparse parity: a *low-degree curriculum* $\rightarrow$ improved sample complexity.
  - Analysis: larger Fourier coefficients on $\{i\}, i \in S$.
  - Generalization: hierarchical parity.
- Generalizing to PCFG & natural languages: *n-gram curriculum*.

ICLR2025

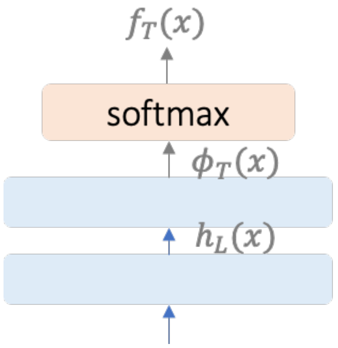
Future Directions

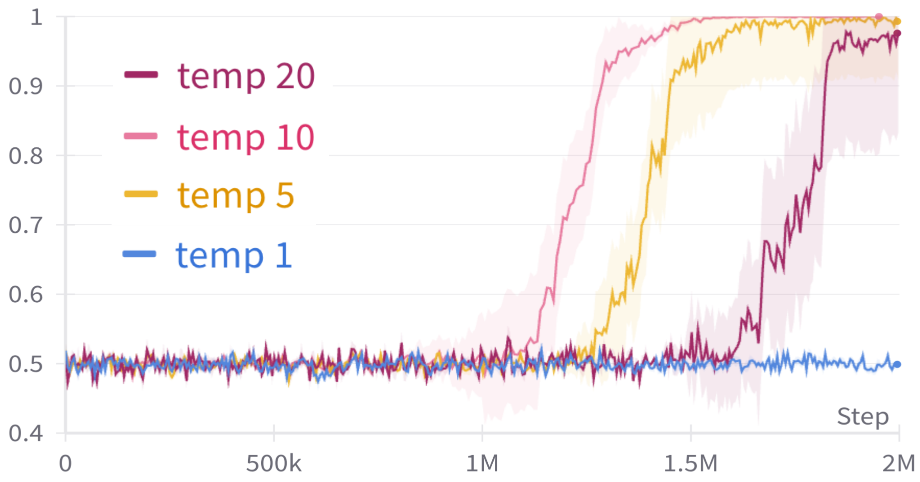
Fewer teachers? For generation? As initialization?

1. Curriculum from a single teacher?

Remove the need to access/store intermediate checkpoints. (e.g. 2-shot for parity)

- A follow-up work: layerwise distillation [Gupta & Karmalkar 25].
- High temperature for the final teacher (a different mechanism?)





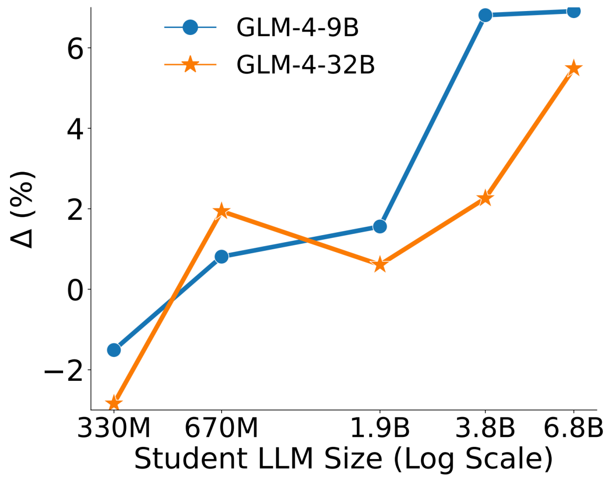
2. Progressive distillation for generations?

Using intermediate teachers for a generative setup (e.g. languages)?

- Not straightforward, based on initial results; though capacity gap does exist.

Method	$\text{BERT}_{\text{base}}$	$\text{BERT}_{\text{large}}$	$\triangle$
Teacher	86.7	88.3	+1.6
$\text{KD}_{10\%/5\%}$ (2015)	81.3	80.8	−0.5
$\text{DynaBERT}_{15\%/5\%}$ (2020)	81.1	79.2	−1.9
$\text{MiniDisc}_{10\%/5\%}$ (2022a)	82.4	82.1	−0.3
$\text{TinyBERT}_{4L;312H}$ (2020)	82.7	82.5	−0.2
$\text{MiniLM}_{3L;384H}$ (2021b)	82.5	82.0	−0.5
$\text{MiniMoE}_{3L;384H}$ (**ours**)	82.6	83.1	+0.5

[Zhang et al. 23]



[Peng et al. 24]

2. Progressive distillation for generations?

Using intermediate teachers for a generative setup (e.g. languages)?

- Not straightforward, based on initial results; though capacity gap does exist.
- Many considerations:
  - Texts or logits (large $|\mathcal{Y}|$)?
  - Format (e.g. CoT)?
  - Teacher-student "alignment/coverage"?

    [Phuong & Lampert 19, Ji & Zhu 20, Harutyunyan et al. 23, Huang et al. 25]

3. Distillation for better initialization?

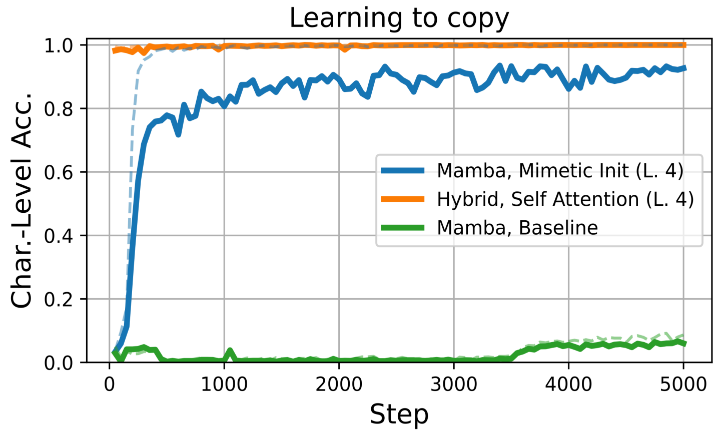
Across model classes, e.g. Transformer to state-space model (SSM) / hybrids.

MODEL	AVG. ↑
Phi-1.5-1.3B	64.9
Phi-Mamba-1.5B	**62.6**
Mamba-1-1.4B	59.7
Mamba-2-1.3B	59.6

[Bick et al. 24]

Model (% Att)	AlpacaEval (win %)
Llama-3-Instruct	$22.60_{1.26}$
Mamba-Llama3 (50%)	$\mathbf{26.69}_{1.31}$
Mamba-Llama3 (25%)	$22.50_{1.26}$
Mamba-Llama3 (12.5%)	$17.93_{1.16}$

[Wang et al. 24]



[Trockman et al. 24]

Goal: progress without massive compute

Train small models better, given big pretrained models?

Distillation for better **efficiency**.

- Training: fewer samples (statistical) / steps (computational).
- Inference: lower cost enabled by performant *small* models.

Progressive distillation accelerates training

(Prior work: *generalization* benefits from full distribution/soft logits)

Intermediate checkpoints provide an implicit curriculum.

- Explains why better teacher $\nrightarrow$ better student ("capacity gap").
- Case study on sparse parity: a *low-degree curriculum* $\rightarrow$ improved sample complexity.
  - Analysis: larger Fourier coefficients on $\{i\}, i \in S$.
  - Generalization: hierarchical parity.
- Generalizing to PCFG & natural languages: *n-gram curriculum*.

ICLR2025

Appendix

Thank you for wanting to know more! :)

Why is distillation helpful?

$\text{loss} = \text{KL}(f_T(x) \| f(x))$

Intuitively: "richer information" ... full distribution vs a sample.

Better generalization: $p^\star(\cdot \mid x)$ leads to a tighter bound [Menon et al. 20].

- Imperfect teacher: bias-variance tradeoff.
- "Teacher-free" via label smoothing $(f_T(x) = (1-\alpha)e_y + \frac{\alpha}{L}\mathbf{1})$ [Yuan et al. 19].

Model	Baseline	Tf-KD$_{reg}$	Normal KD [Teacher]
MobileNetV2	68.38	70.88 (**+2.50**)	71.05 (**+2.67**) [ResNet18]
ShuffleNetV2	70.34	72.09 (**+1.75**)	72.05 (**+1.71**) [ResNet18]
ResNet18	75.87	77.36 (**+1.49**)	77.19 (**+1.32**) [ResNet50]
GoogLeNet	78.15	79.22 (**+1.07**)	78.84 (**+0.99**) [ResNeXt29]

Signals: Fourier coefficients on x_i, $x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- How: population gradient at initialization [Edelman et al. 22].

  Consider a single neuron $w \in \mathbb{R}^d$:

$f(x) = \sigma(w^\top x + b)$

$l(y, y') = -yy'$

$$-\widehat{\mathrm{LTF}}_{S'} \leftarrow g_i := (\nabla_w \mathbb{E}_x[l(y, f(x; w))])_i = -\nabla_w \mathbb{E}_x[1[w^\top x + b \geq 0] \cdot y x_i]$$

$$= -\mathbb{E}_x[1[w^\top x + b \geq 0] \cdot (\prod_{j \in S} x_j) \cdot x_i]$$

LTF

$\chi_{S'}$,

$S' = S \backslash \{i\}$ (if $i \in S$) or $S \cup \{i\}$ (if $i \notin S$)

Fact: $|\widehat{\mathrm{LTF}}_{S_1}| > |\widehat{\mathrm{LTF}}_{S_2}|$

for odd $|S_1|, |S_2|$ s.t. $|S_1| < |S_2|$.

Signals: Fourier coefficients on $x_i, x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- How: population gradient at initialization [Edelman et al. 22].

  Consider a single neuron $w \in \mathbb{R}^d$:

$$f(x) = \sigma(w^\top x + b)$$
$$l(y, y') = -yy'$$

$$-\widehat{\mathrm{LTF}}_{S'} \leftarrow g_i := (\nabla_w \mathbb{E}_x[l(y, f(x; w))])_i = -\nabla_w \mathbb{E}_x[1[w^\top x + b \geq 0] \cdot y x_i]$$

$$= -\mathbb{E}_x[1[w^\top x + b \geq 0] \cdot (\prod_{j \in S} x_j) \cdot x_i]$$

(Fourier gap)

$$|g_i| \geq |g_j| + \gamma_k, i \in S, j \notin S.$$

large gradients → support

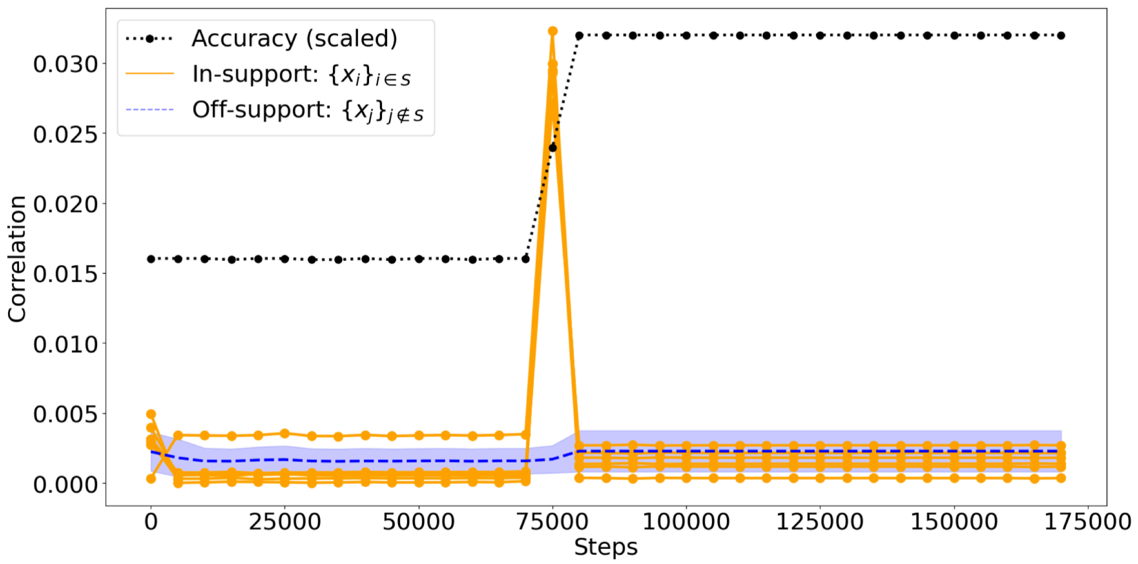
Fact: $|\widehat{\mathrm{LTF}}_{S_1}| > |\widehat{\mathrm{LTF}}_{S_2}|$

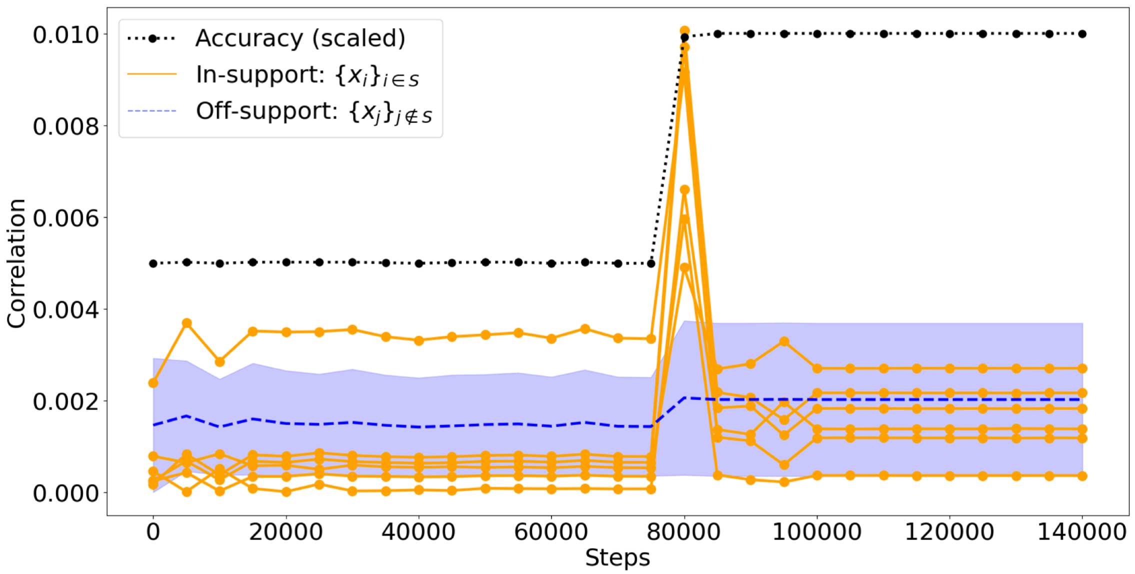
for odd $|S_1|, |S_2|$ s.t. $|S_1| < |S_2|$.

$\chi_{S'}$, $S' = S \backslash \{i\}$ (if $i \in S$) or $S \cup \{i\}$ (if $i \notin S$)

Transformer on sparse parity

1. **Implicit curriculum** emerges: Higher $\hat{f}_{\{i\}}$ for $i \in S$.

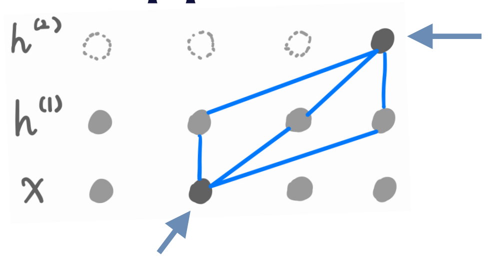


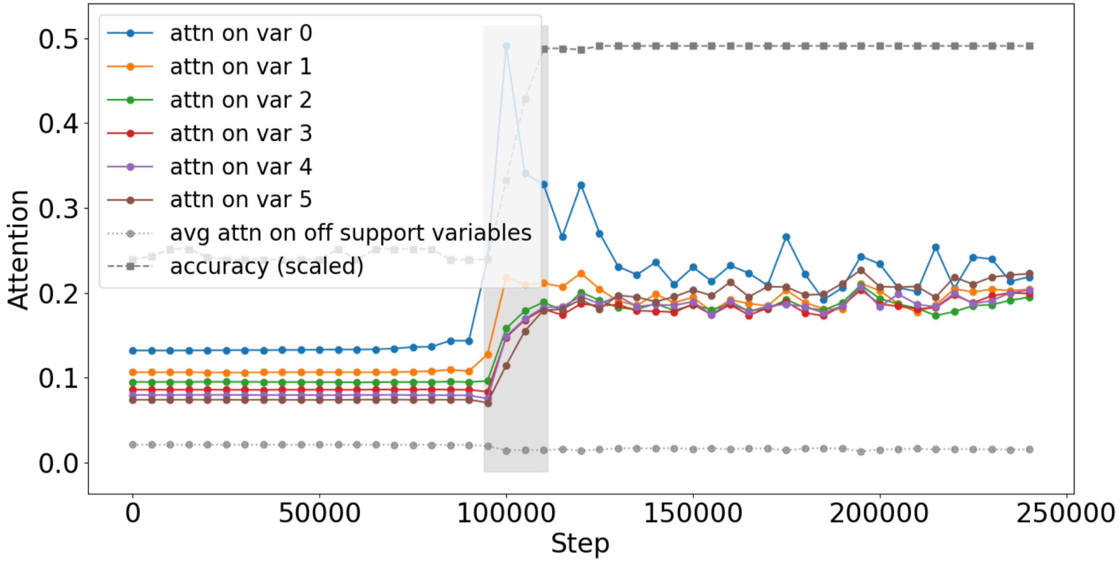


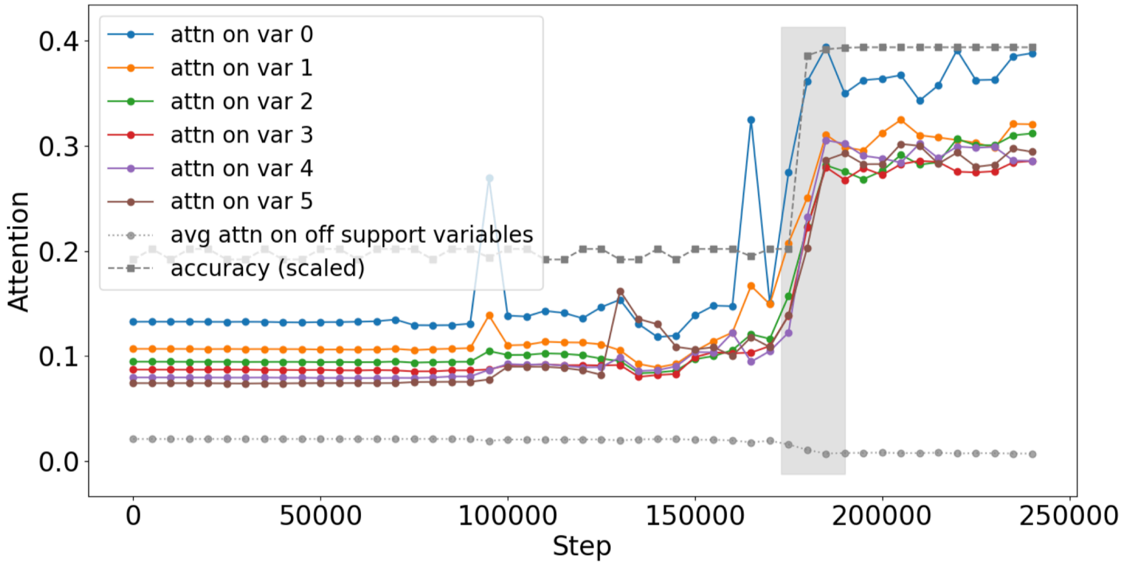
Transformer on sparse parity

2. **True support** is learned: more attention weights on in-support coordinates.

sum of length-2 paths

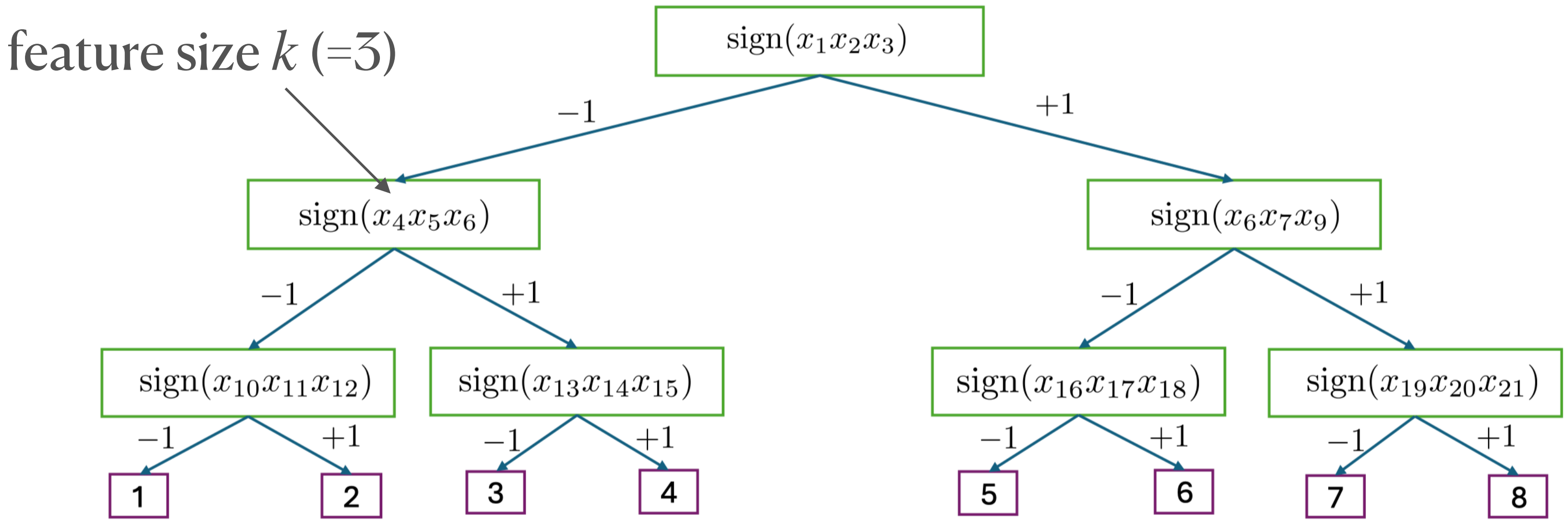






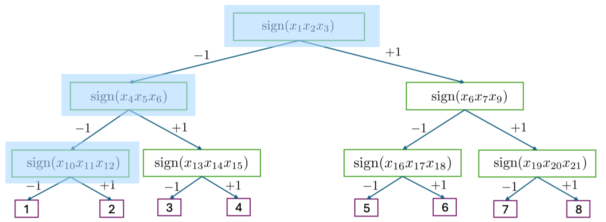
Beyond sparse parity — a hierarchical task

Hierarchical parity … depth-D $\rightarrow$ 2^D-way classification.

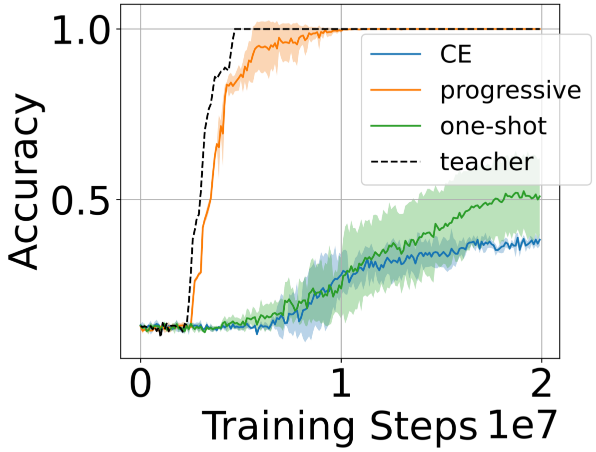


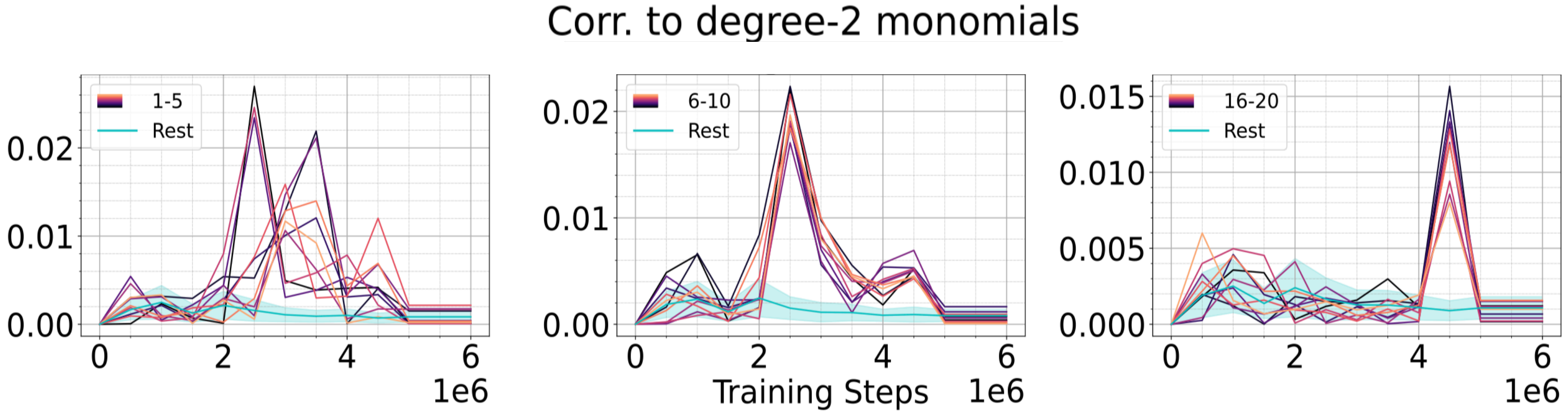
Beyond sparse parity — a hierarchical task

Hierarchical parity ... depth-D $\rightarrow$ 2^D-way classification.



- Results on $d = 100, D = 3, k = 5$:





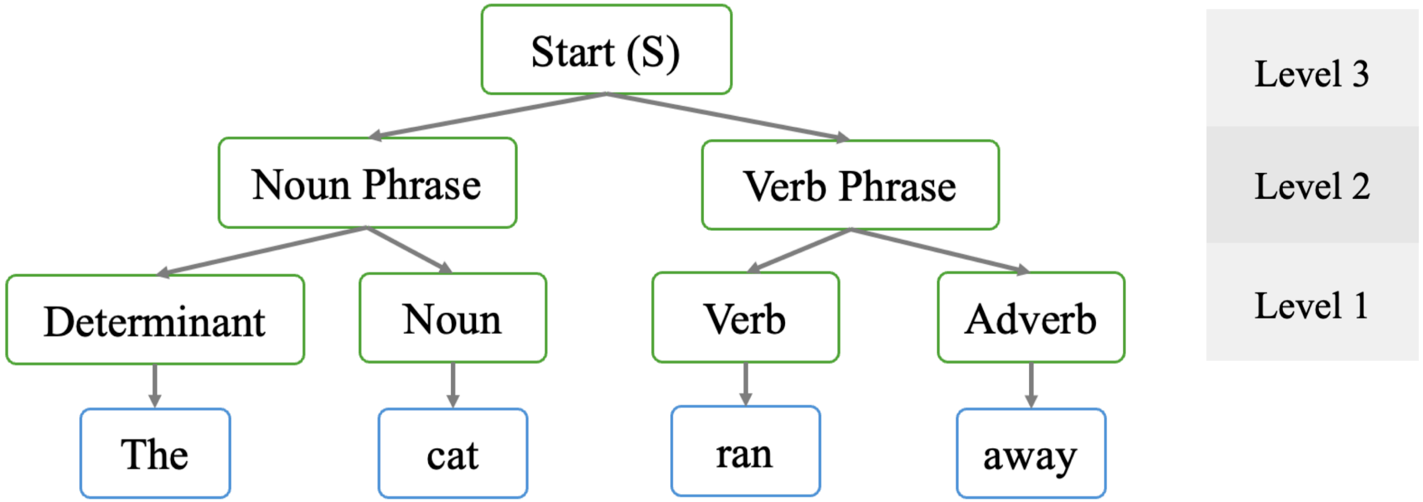
Learning at diff speed → need multiple teachers.

Beyond sparse parity — formal languages

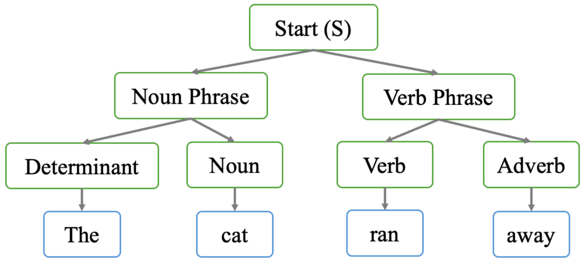
Data: **PCFG** (probabilistic context-free grammar) [Allen-Zhu & Li 23]

(non-terminals)

(vocab)



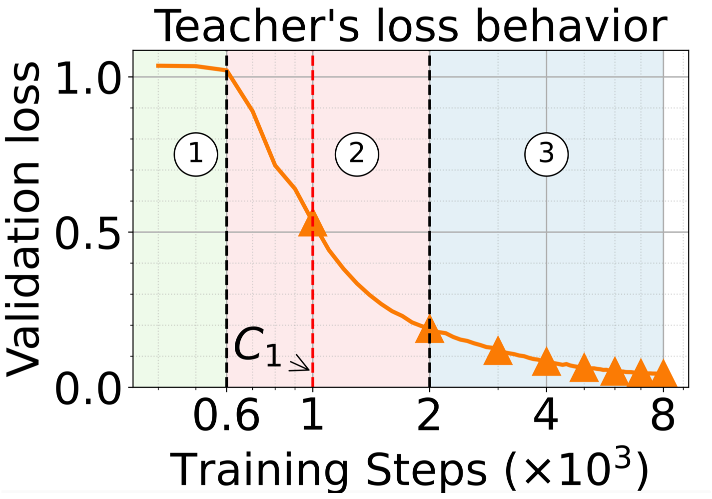
Beyond sparse parity — PCFG

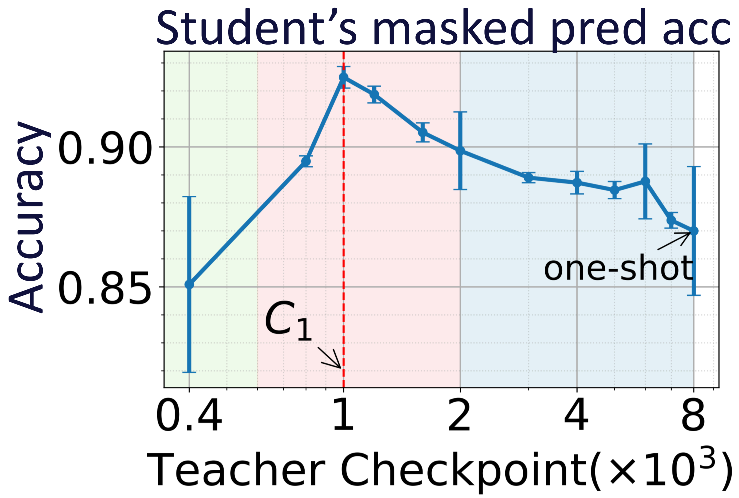


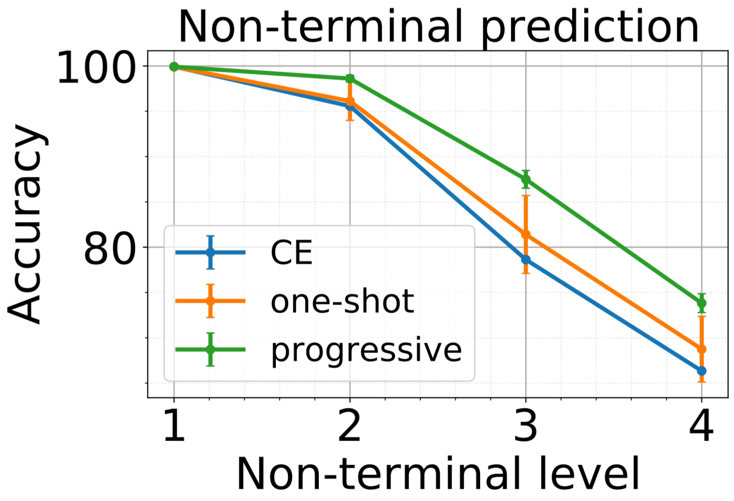
Task: masked prediction → optimal: following the tree hierarchy [Zhao et al. 23].

a quality measure

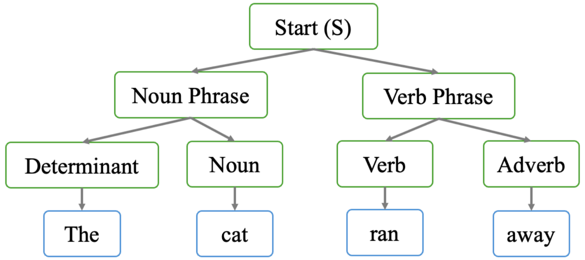
An implicit curriculum exists. *… what is it?*







n-gram curriculum for PCFG



n-grams with an increasing n. (e.g. $n = 3$: cat ran away, cat danced away, cat jumped away, …)

- Smaller n (more local/lower sensitivity) is easier [Abbe et al. 23,24; Vasudeva et al. 24].

2 measures for the dependency on n-grams:

- $M_{\text{robust}} = \text{TV}\big(p(x_{\backslash\{i\}}),\, p(x_{\backslash n\text{-gram}(i)})\big)$ The cat ____ _?_ _____ after hearing…
  - "All but n-gram": smaller → the prediction depends less on n-gram.
- $M_{\text{close}} = \text{TV}\big(p(x_{\backslash\{i\}}),\, p(x_{n\text{-gram}(i)\backslash\{i\}})\big)$ ____ ____ ran _?_ away _____ ______…
  - "Only n-gram": smaller → the prediction is closer to a n-gram model.