arxiv:2410.05464

arxiv:2511.02833

Learning from the Right Teacher in Knowledge Distillation

Bingbin Liu

Kempner Institute,
Harvard University

Abhishek
Panigrahi

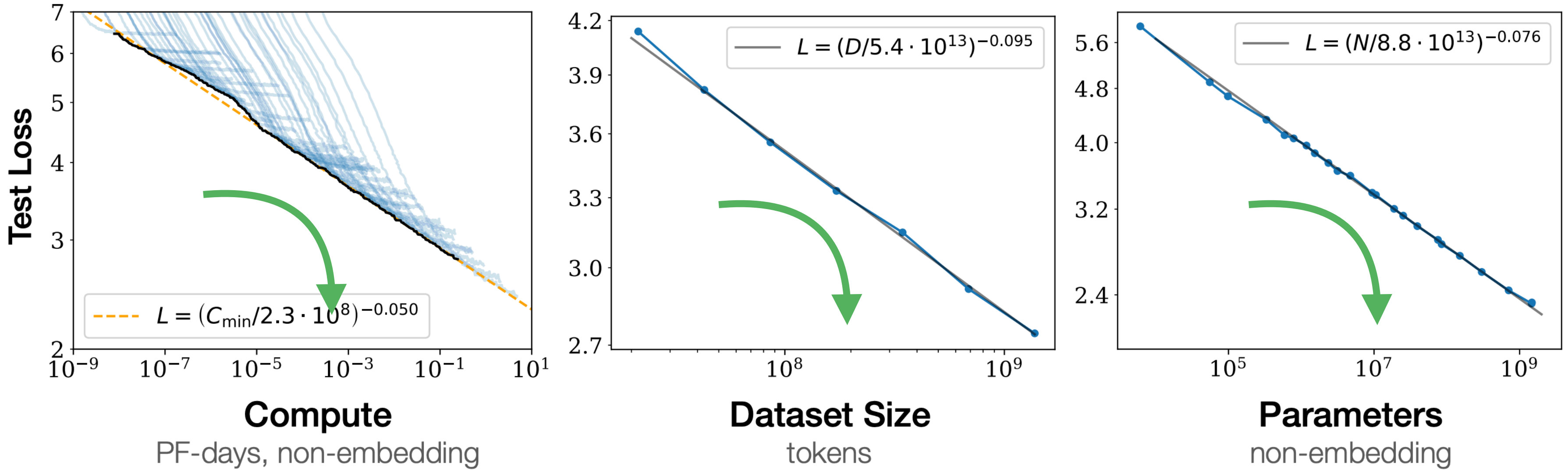
Sadhika
Malladi

Sham
Kakade

Andrej
Risteski

Surbhi
Goel

Progress at a small scale?



[Hoffmann et al. 22]

Progress at a small scale?

e.g. better performance at a small model size?

Why this matters:

1. Performance = resource * (performance / resource).

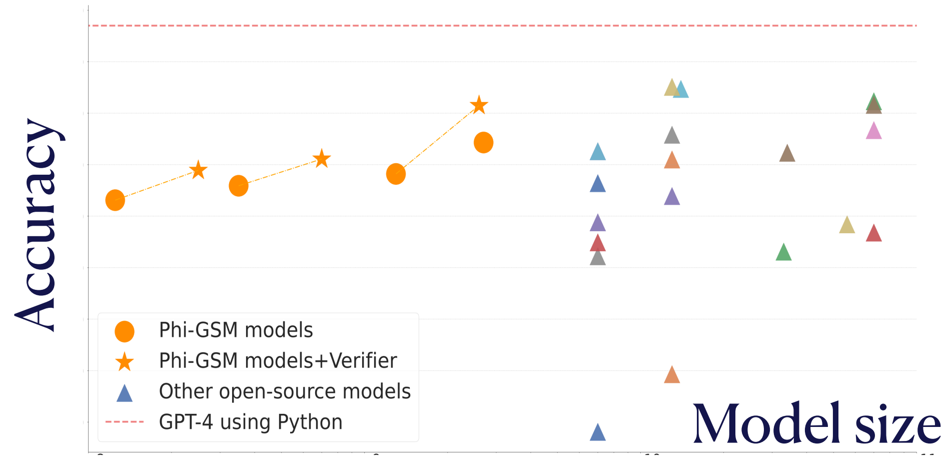
2. Cost ... training and inference cost $$$.

3. Accessibility ... research and usage.

Progress at a small scale? ... with distillation

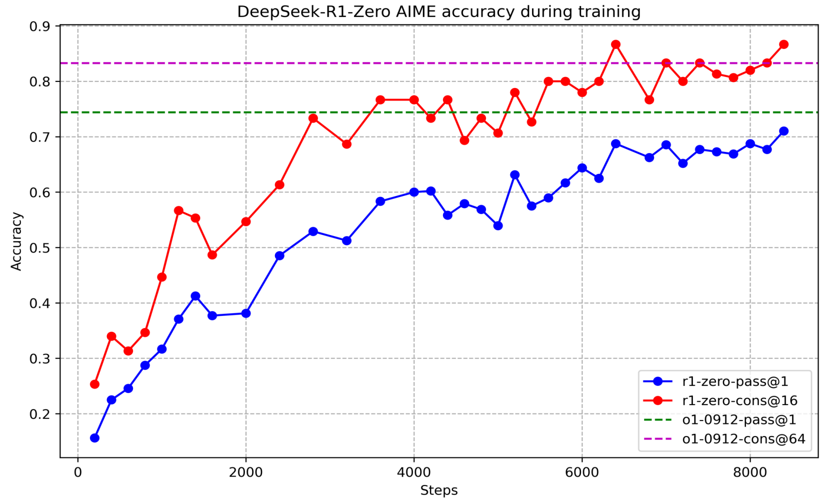
e.g. better performance at a small model size? Possible!

Data



[Liu et al. 23]

Algorithm



[DeepSeek-R1]

This talk: knowledge distillation

Better small models, by leveraging powerful pretrained models.

- **Faster 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]

This talk: knowledge distillation

Better small models, by leveraging powerful pretrained models.

- **Better inference**: performant small models; "model compression".

or: 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]

This talk: knowledge distillation

Better small models, by leveraging powerful pretrained models.

??

- **Better inference**: performant small models; "model compression".

or: quantization, pruning.

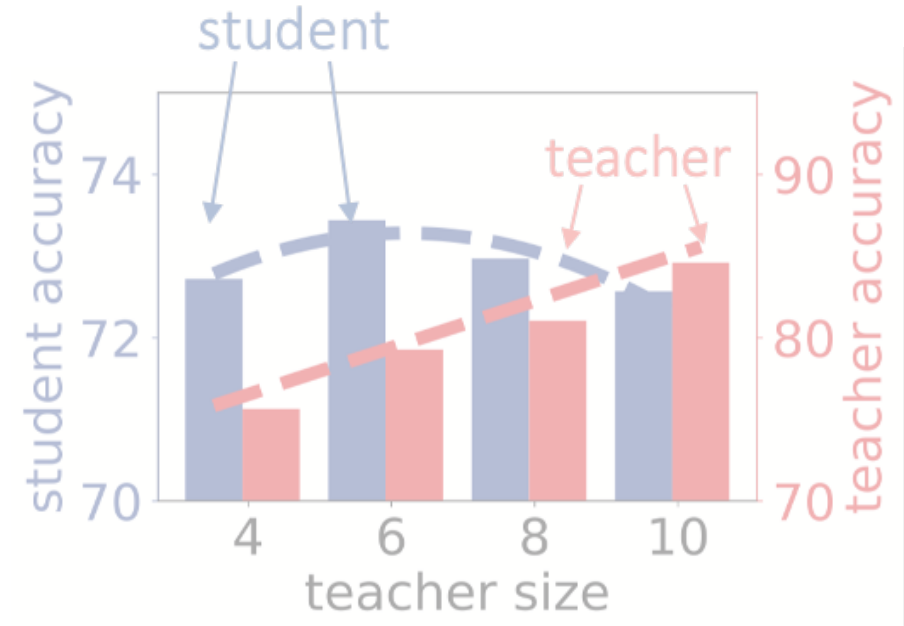
Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

[DeepSeek R1 report]

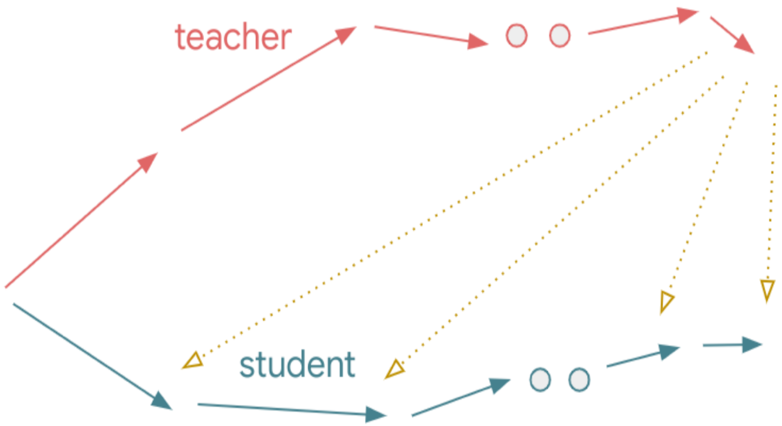
Stronger teacher ↛ better student

"capacity gap"

(due to differences in size / training steps)



[Mirzadeh et al. 19]



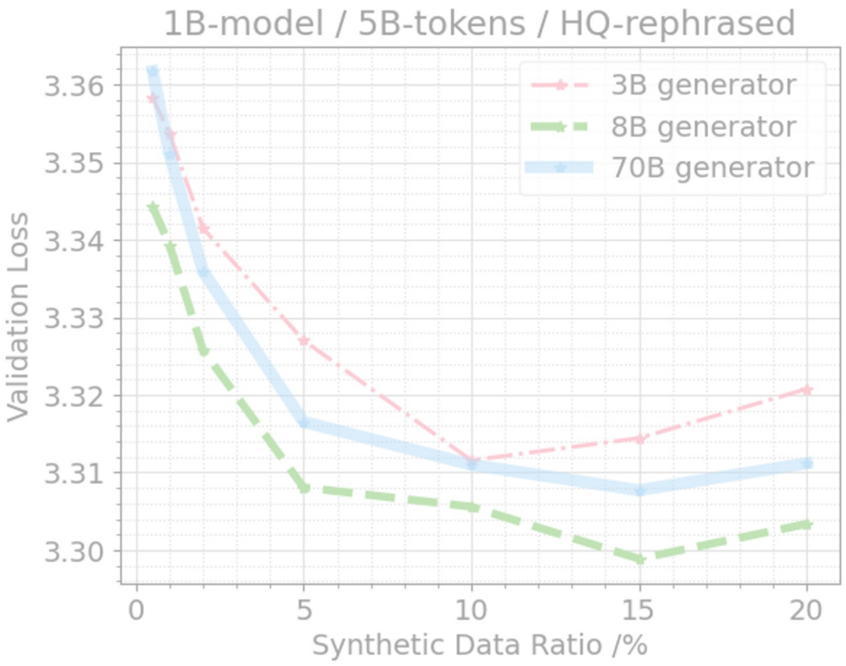
[Harutyunyan et al. 23]

Stronger teacher $\nrightarrow$ better student

"capacity gap"
(due to differences in size / training steps)

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]



[Kang et al. 25]

Distilling from the *right* teacher

Progressive distillation: implicit curricula from *intermediate checkpoints*.

- Case study on sparse parity: improved sample complexity.
- Empirically verified more broadly.

GRACE for teacher selection: scoring teachers for LLM post-training.

- Indicative of the student's performance.
- Guide design choices.

Part 0: knowledge distillation background

What is knowledge distillation?

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

- Classification with cross-entropy loss: $f(x) \in \Delta^{k-1}, y \in [k]$.
(per-token for LLM)

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

Distillation from f_T: $L_D(f(x), f_T(x)) = \mathrm{KL}(f_T(x) \| f(x))$.

reverse KL, l_2, etc.

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

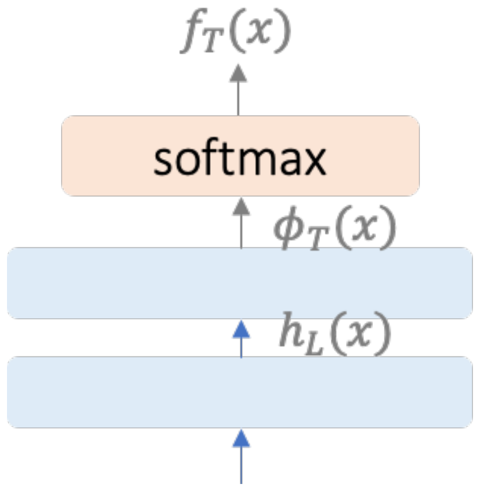
How to knowledge distill?

Mimic the teacher's outputs or intermediate activations.

- Post-softmax output $[f(x)]_i \propto \exp(\ \tau^{-1} \cdot [\phi(x)]_i\ )$.

  (inverse) temperature

  - Intermediate activation: need to match dimension.



- For LLMs / (probabilistic) generative models:
  - Per-token logits (pre/post-softmax).
  - Samples: e.g. synthetic data.

How to knowledge distill?

Various scenarios

(by capabilities)

- Big/strong teacher → small/weak student.
- Same-sized teacher & student: self-distillation.
- Small/weak teacher → big/strong student (weak-to-strong).

(by quantities)

- An ensemble of teachers → a single student.
- A single teachers → a series of students.

Why is distillation helpful?

Intuition: "**richer information**" ... e.g. class relation, per-sample weighting.

- An ideal teacher: $f_T(x) = p^\star(y \mid x)$, i.e. providing the full label distribution.

more informative than the data
i.e. a single $y \sim p(\cdot \mid x)$.

Soft labels / $p^\star(y \mid x)$ lead to **better generalization** [Menon et al. 20].

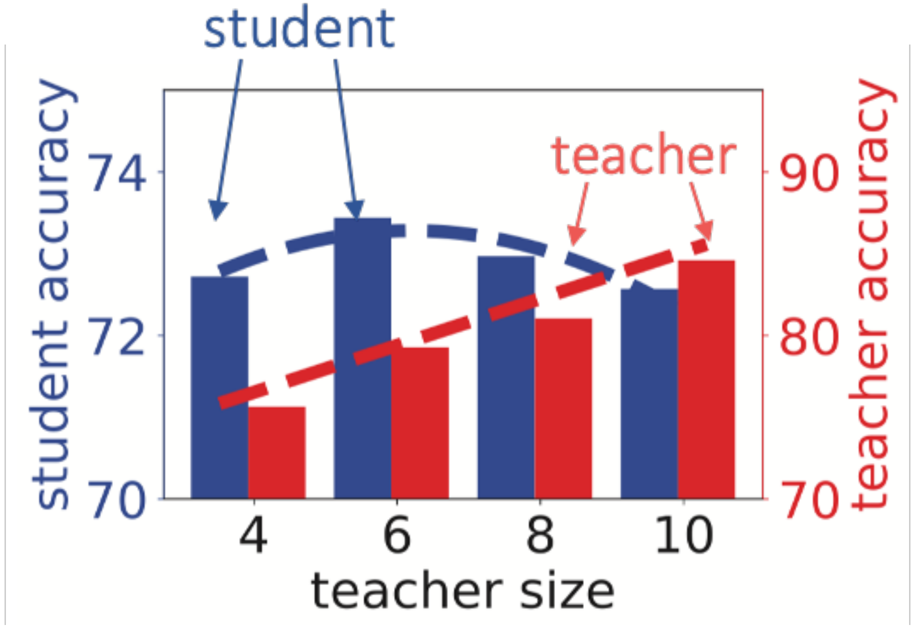
- Part 1: effect on *optimization*.

Part 1: Progressive distillation

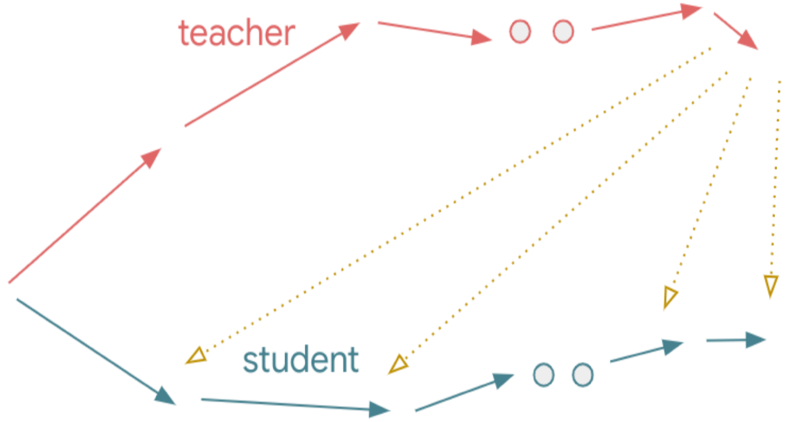
Implicit curriculum from intermediate checkpoints

Recall: stronger teacher ↛ better student

"capacity gap"
(due to differences in size / training steps)



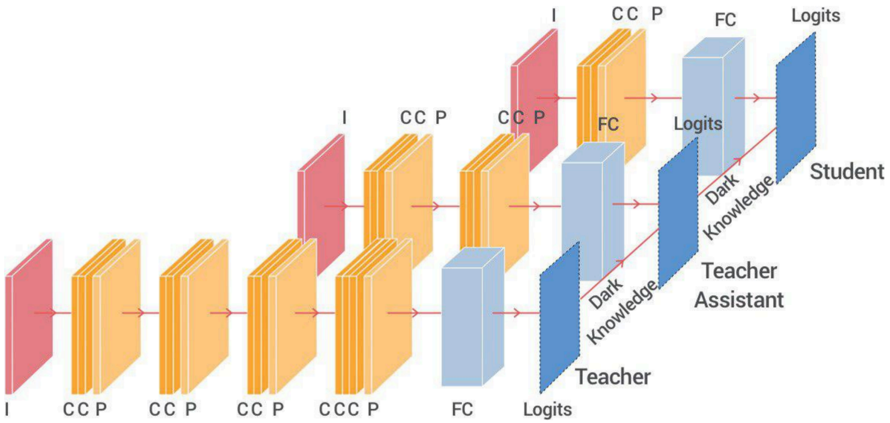
[Mirzadeh et al. 19]



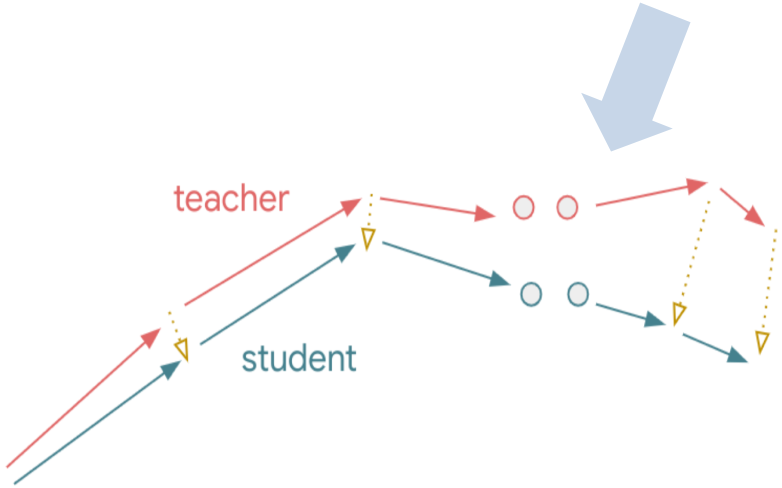
[Harutyunyan et al. 23]

Recall: stronger teacher ↛ better student

Closing the "capacity gap".
(intermediate sizes / training steps)



[Mirzadeh et al. 19]



[Harutyunyan et al. 23]

Why intermediate teachers help?

Prior work: better generalization (upper) bounds [Harutyunyan et al. 23].

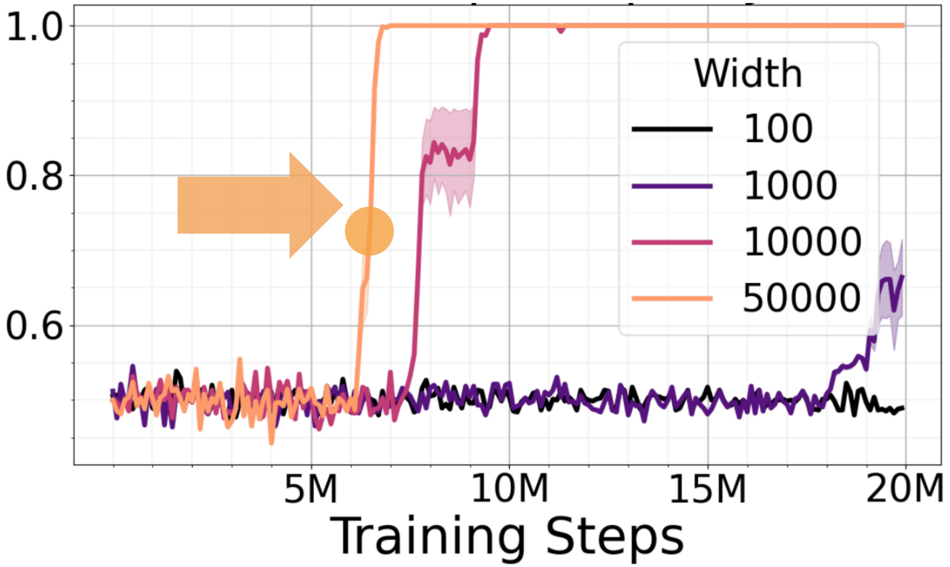
Our work: an **optimization** perspective.

- Intuition: using teacher's trajectory to guide the student's optimization.
- Case study: **sparse parity** … prior theory fails to explain the gain.
- Empirical validation on more realistic settings (PCFG and natural languages).

Case study: sparse parity

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

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

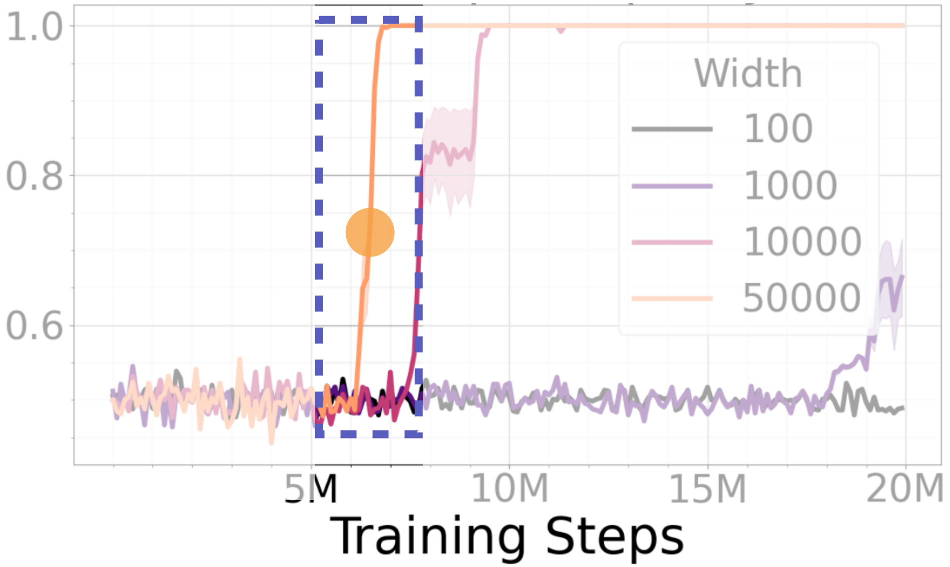


Case study: sparse parity

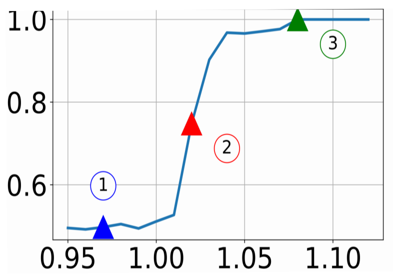
$$x = 1\ \text{-}1\ \underbrace{\text{-}1\ 1\ \ \text{-}1}_{S}\ 1\ 1\ 1\ \text{-}1\ 1 \in \mathbb{R}^d \rightarrow y = \prod_{i \in S} x_i = 1$$

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

which ones?



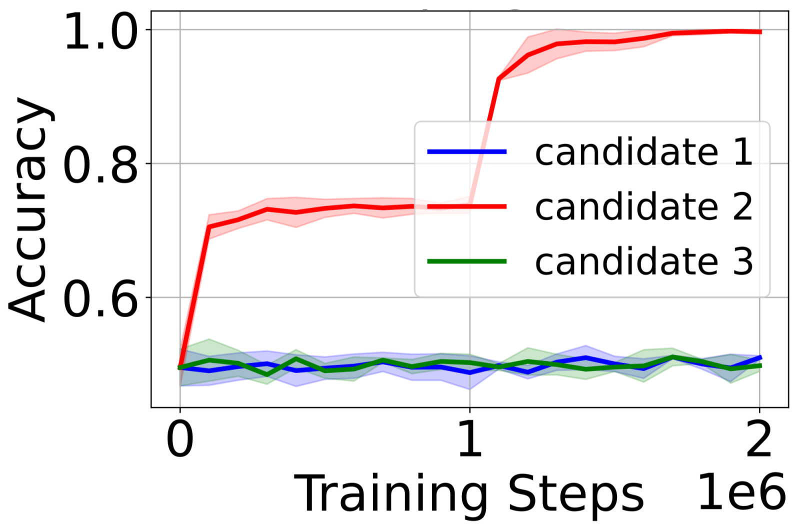
Signals from intermediate teachers



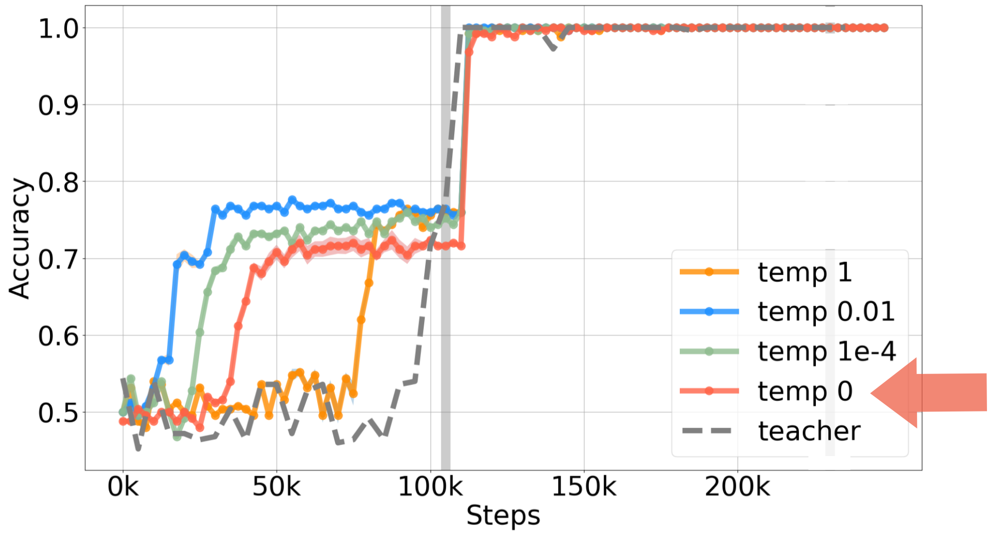
Setup: learning from 1 intermediate teacher + the final teacher.

- Choices: before / during / after the phase transition.

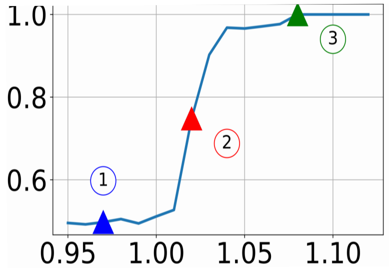
1. Lead to a better student.



Not because of "soft labels".



Signals from intermediate teachers

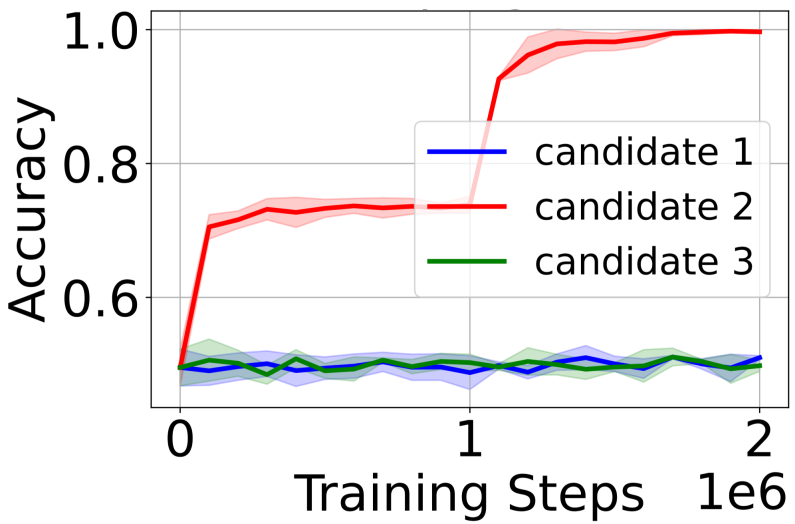


Setup: learning from **1 intermediate teacher** + the final teacher.

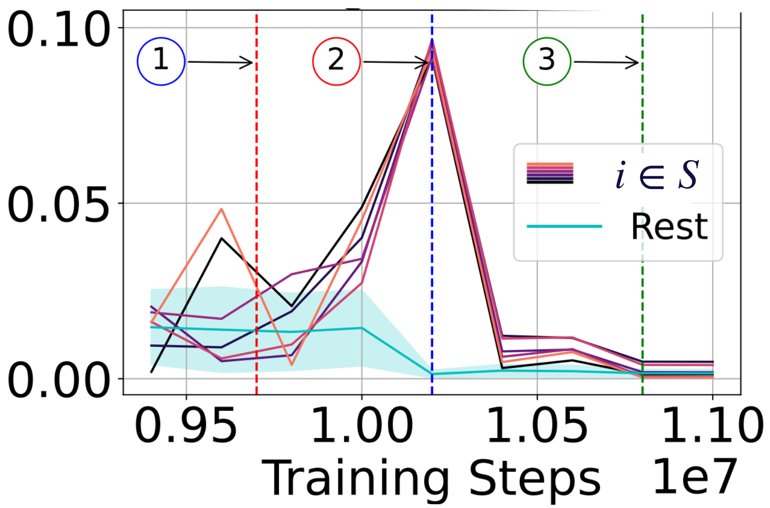
- Choices: before / during / after the phase transition.

Implicit curriculum

1. Lead to a better student.



2. Providing "**extra signals**".



Implicit curriculum helps with optimization

- Case study: sparse parity: speedup from “extra training signals.”
  - What are the signals? … Fourier coefficients.
  - Why are they helpful? … sample complexity.
  - How do they emerge in the teacher? … initial population gradient.
- Progressive distillation & empirical validation

Training setup

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

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

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

- Teacher: 2-phase: initial large batch, followed by online SGD.
- Student: 2-shot distillation, from the end of each phase.

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

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

- Natural for sparse parity: $y_S = \chi_S$.
- Fourier coefficients = projections onto the basis:

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

Signals: Fourier coefficients on x_i, $i \in S$

Implicit curriculum via $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in S$).

Learning from $y = \chi_S(x) \rightarrow \Omega(d^k)$ samples (recall: $|S| = k$).

Learning from $\sum_{i \in S} \chi_{\{i\}} \rightarrow \Omega(d)$ samples.

$\tilde{\Theta}_{k,\epsilon}(d^2)$ for student's 2-shot distillation.

- Why: sample complexity to learn $\chi_{\tilde{S}}$: $\Omega(d^{|\tilde{S}|})$ (SQ lower bound).

$\rightarrow$ Fewer samples for learning lower-degree monomials [Edelman et al. 22, Abbe et al. 23].

Signals: Fourier coefficients on x_i, $i \in S$

Implicit curriculum via $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in S$).

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

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

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

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

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

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 $i \notin S$)

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

Implicit curriculum via $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in S$).

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

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

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

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

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

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

(Fourier gap)

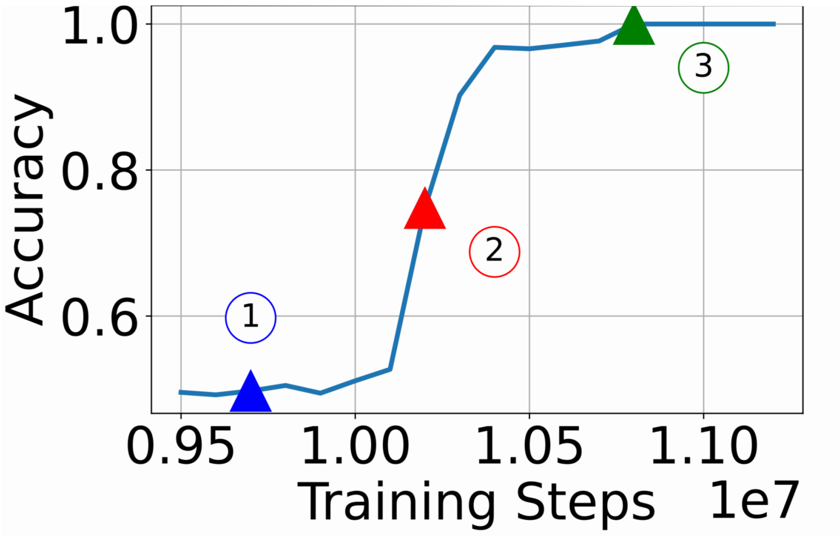
$$|g_i| \geq |g_j| + \gamma_k, i \in S, j \notin S.$$

large gradients → support

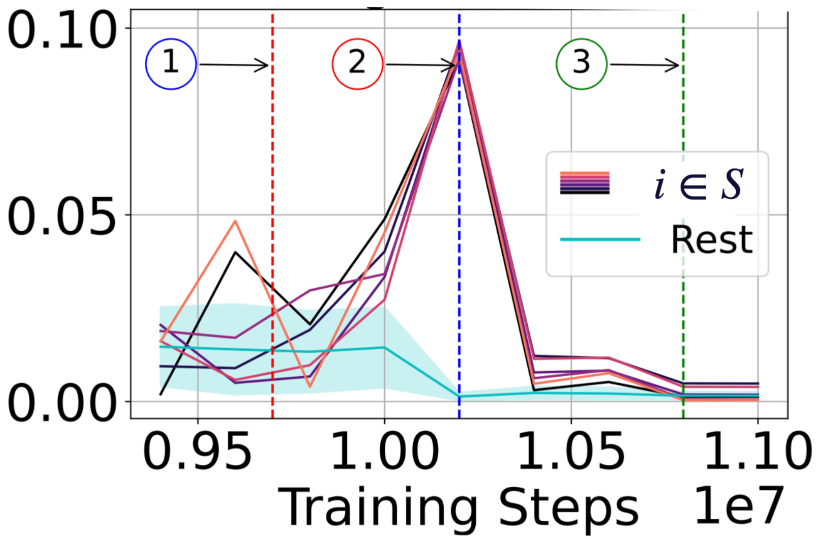
$\chi_{S'}$, $S' = S \backslash \{i\}$ (if $i \in S$) or $S \cup \{i\}$ (if $i \notin S$)

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

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



larger $|\chi_{\{i\}}|$ for $i \in S$



Implicit curriculum helps with optimization

- Sparse parity: faster feature learning from "extra training signals."
  - **What** the curriculum are: Larger Fourier coeffs on x_i, $i \in [S]$.
  - **Why** they are helpful: sample complexity $\Omega(d^k) \to \tilde{\Theta}(d^2)$.
  - **How** they emerge: Initial population gradient reveals the support.

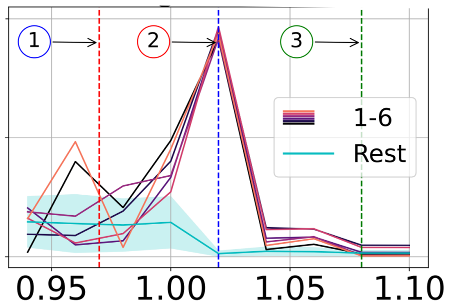
Implicit curriculum: a helpful decomposition.

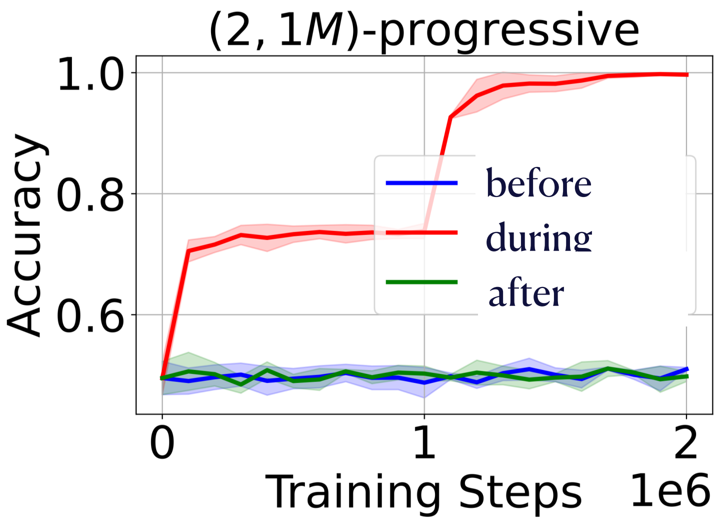
Next: progressive distillation & empirical validation

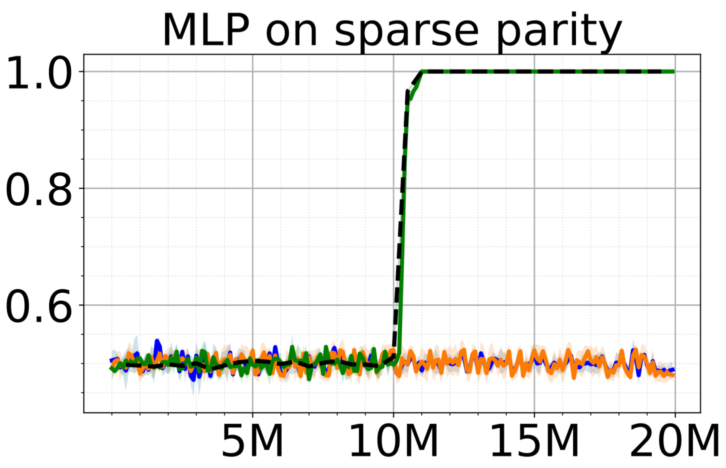
Progressive distillation

Distilling from checkpoints at certain intervals.

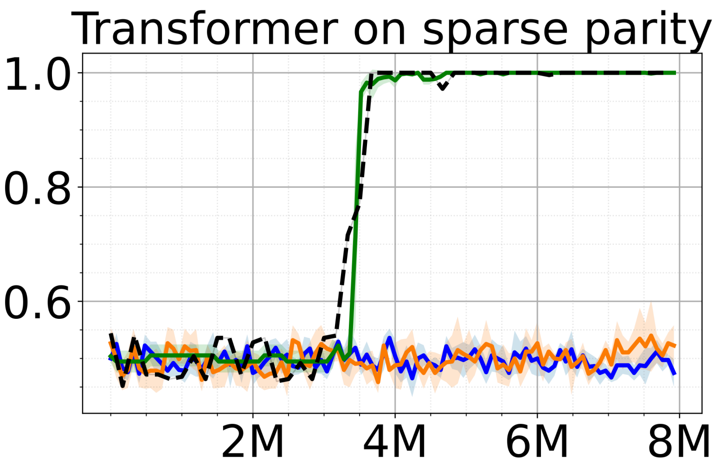
larger $|\chi_{\{i\}}|$ for $i \in S$





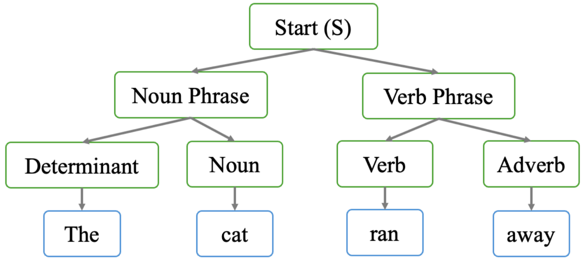


at 100k intervals



at 10k intervals

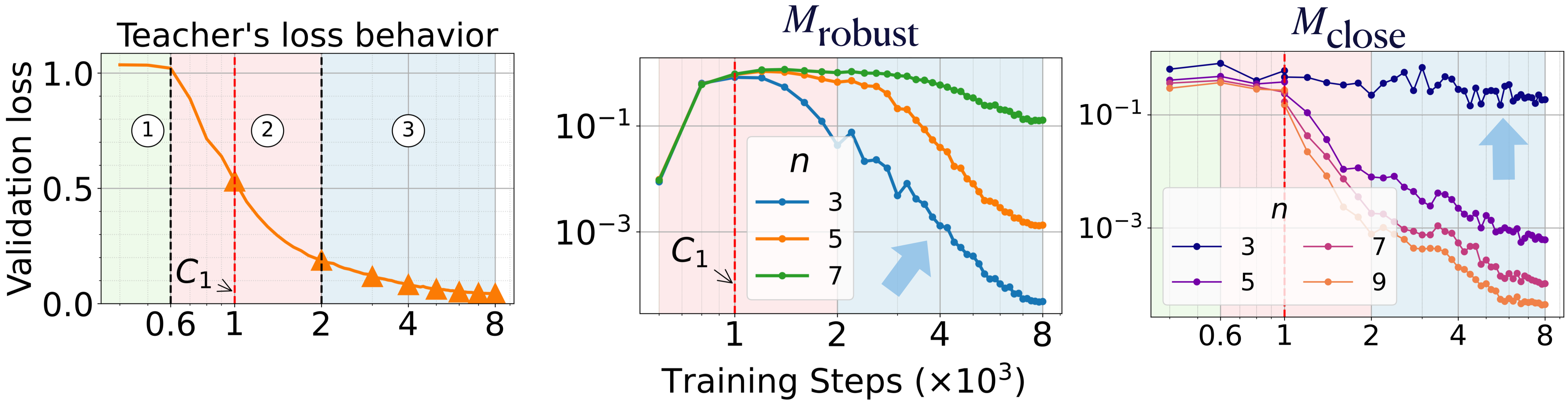
Beyond sparse parity



Task: Masked prediction on formal (PCFG) and natural (Wiki/Books) languages.

Implicit curriculum: *n*-grams with an increasing *n*.

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



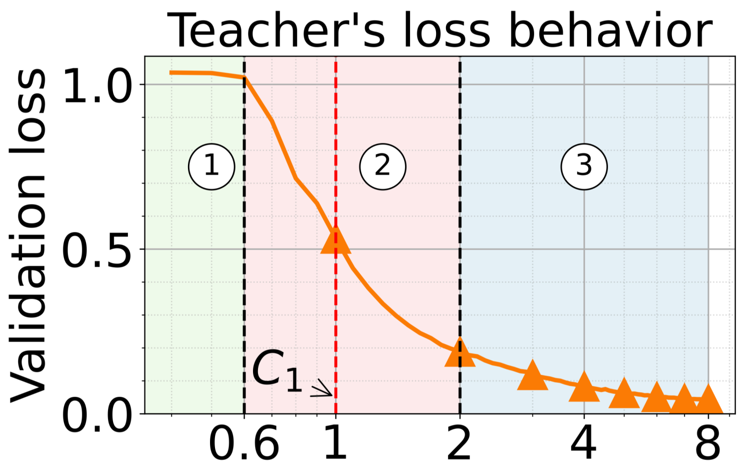
Beyond sparse parity

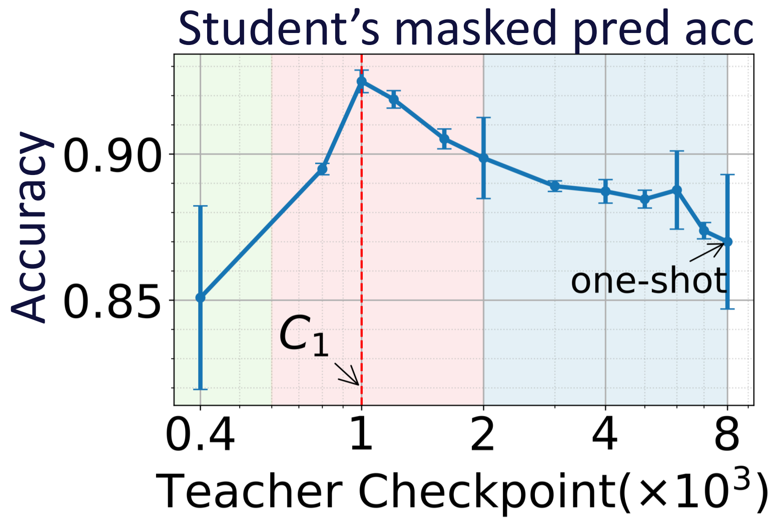
Task: Masked prediction on formal (PCFG) and natural (Wiki/Books) languages.

Implicit curriculum: n-grams with an increasing n.

Results: "phase-transition" checkpoints are the best teachers.

- PCFG:





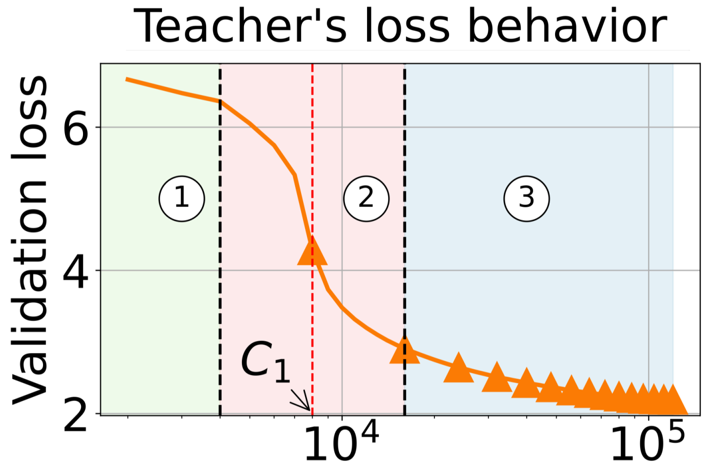
Beyond sparse parity

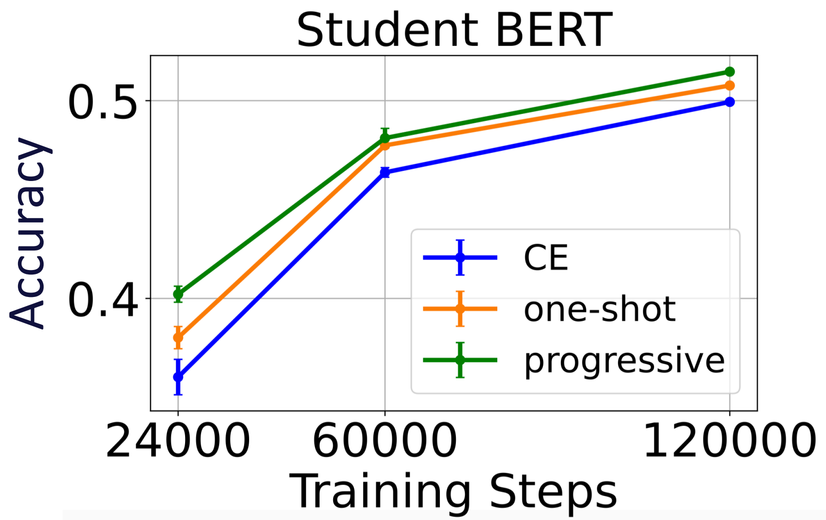
Task: Masked prediction on formal (PCFG) and natural (Wiki/Books) languages.

Implicit curriculum: *n*-grams with an increasing *n*.

Results: “phase-transition” checkpoints are the best teachers.

- Wiki/Books:



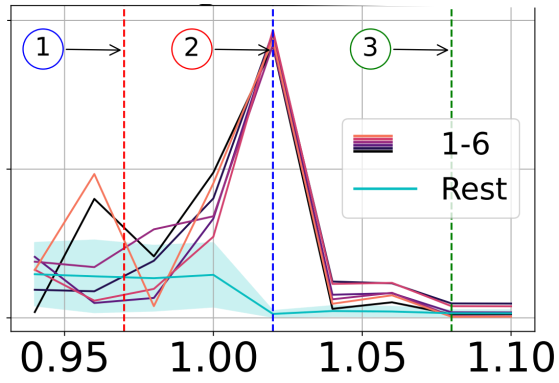


Part 1: right teacher for faster training

Progressive distillation induces an implicit curriculum that accelerates optimization.

- Sparse parity: a *low-degree curriculum* → improved sample complexity.
  - Analysis: larger Fourier coefficients on $\{i\}, i \in S$.
  - Generalization: hierarchical parity.
- PCFG & natural languages: *n-gram curriculum*.

larger $|\chi_{\{i\}}|$ for $i \in S$



Part 2: The GRACE Score

Teacher selection for LLM post-training

Knowledge distillation for LLMs

1. Efficient post-training.
   (can be better than RL)

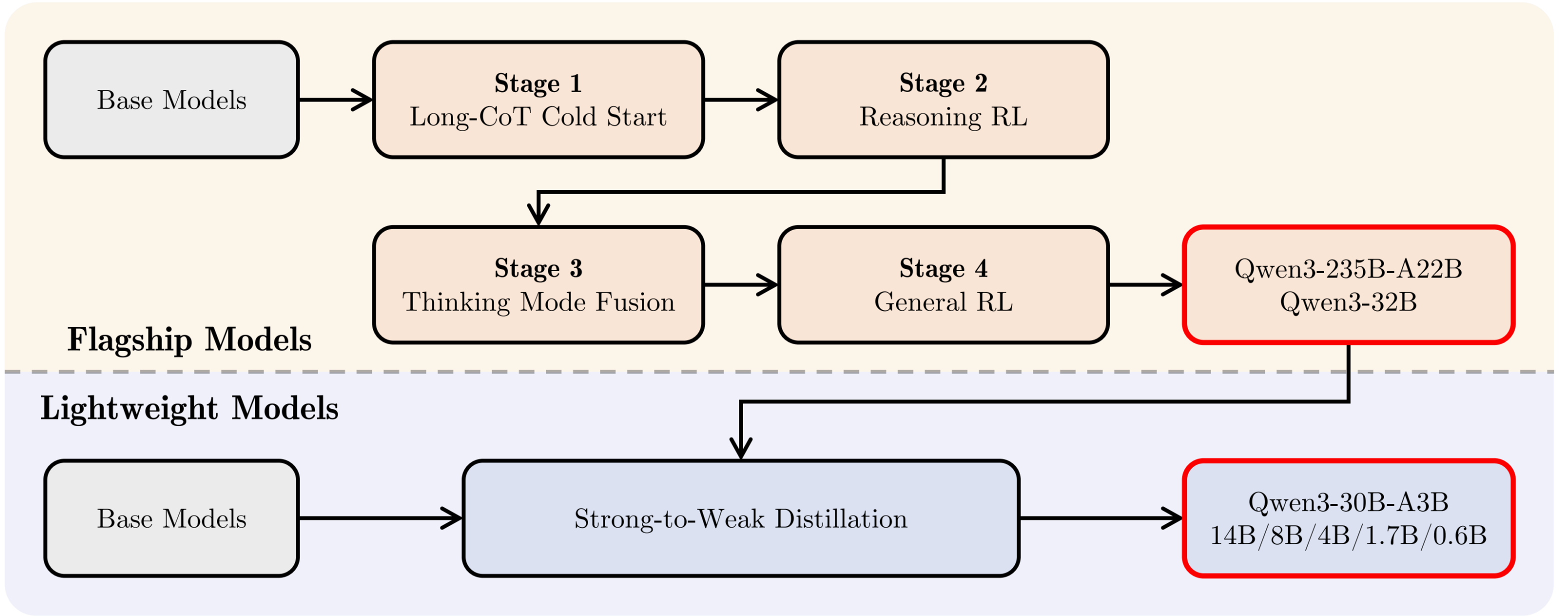
Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

[DeepSeek R1 report]

Knowledge distillation for LLMs

1. Efficient post-training.

2. Make the model RL-able.



[Qwen3 report]

Knowledge distillation for LLMs

So many choices…!

(Recall: capacity gap)

Qwen

Microsoft Phi-4

Gemma

Ai2 OLMo Open Language Model

deepseek

KIMI

GRACE: a score for LLM teacher selection

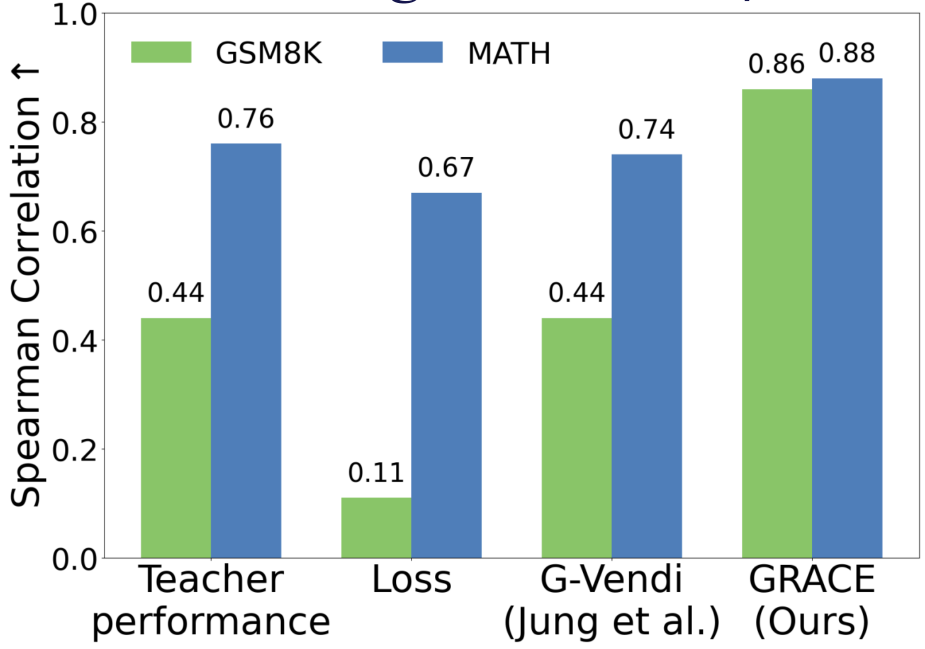
(GRAdient Cross-validation Evaluation)

1) Indicative of the student's performance;

2) Informing distillation practices.

- generation temperature?
- teacher under a constrained size?
- teacher within a family?

strong correlation ↑



Designing GRACE

Commonly used for data selection.

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

- Computationally light-weight ... compared to training.
- Black-box access suffices ... e.g. API access.
- Applicable across model families ... no tokenizer issues.

Designing GRACE

GRACE: gradient norm, weighted by the spectrum of normalized gradients.

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

For a (small) dataset D, compute gradients $G \in \mathbb{R}^{N \times d}$ with rows $\{g_x\}$.

- Gradient covariance $\Sigma := G^\top G \in \mathbb{R}^{d \times d}$.
- Take a partition $D_1 \cup D_2 = D$; compute Σ_1, and $\tilde{\Sigma}_2$.

on normalized gradients

leave-one-out conditional mutual info $\lesssim$ GRACE.

Designing GRACE

GRACE: gradient norm, weighted by the spectrum of normalized gradients.

$$\textbf{GRACE}(D) := \hat{\mathbb{E}}_{D_1 \cup D_2 = D}\left[\mathrm{Tr}\left(\Sigma_1 \tilde{\Sigma}_2^{-1}\right)\right]$$

For a (small) dataset D, compute gradients $G \in \mathbb{R}^{N\times d}$ with rows $\{g_x\}$.

- Gradient covariance $\Sigma := G^\top G \in \mathbb{R}^{d\times d}$.
- Take a partition $D_1 \cup D_2 = D$; compute Σ_1, and $\tilde{\Sigma}_2$.

on normalized gradients

leave-one-out conditional mutual info $\lesssim$ GRACE.

Designing GRACE

GRACE: gradient norm, weighted by the spectrum of normalized gradients.

$$\textbf{GRACE}(D) := \hat{\mathbb{E}}_{D_1 \cup D_2 = D}\left[\mathrm{Tr}\underbrace{\left(\Sigma_1 \tilde{\Sigma}_2^{-1}\right)}\right]$$

$\sum_{i\in[d]} \lambda_i^{-1}\mathbb{E}_{x\sim D_1}\left(\langle v_i, g_x\rangle^2\right),\ \{\lambda_i, v_i\}$ on D_2.

scale with gradient norm

- Norm only: G-Norm $:= \mathbb{E}_{x\sim D}||g_x||^2$.
- Spectrum only: G-Vendi (Jung et al. 25) $:= -\sum_{i\in[d]} \lambda_i \log \lambda_i$.

Teacher selection for math reasoning

(lots of teachers available)

- Datasets: GSM8K (left), MATH (right).

Q: Jin earns $12 an hour. She did 50min of work today. How much did she earn?

A: Jin earns 12/60 = 0.2 per minute. For 50 minutes, she earned 0.2 x 50 = 10.

Q: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles.
How many different groups of two marbles can Tom choose?

A: (step-by-step solutions)

Teacher selection for math reasoning

(lots of teachers available)

- Datasets: GSM8K (left), MATH (right).
- Students: Llama-1B / OLMo-1B / Gemma-2B (GSM); Llama-3B (MATH).
  - Performance: **average-at-16**.
- Teachers: 15 models in Gemma, Llama, OLMo, Phi, Qwen (Math), at temperature [0.3,1].
  - Quality: **correlation** to & **regret** of average-at-16.

GRACE for LLM teacher selection

(GRAdient Cross-validation Evaluation)

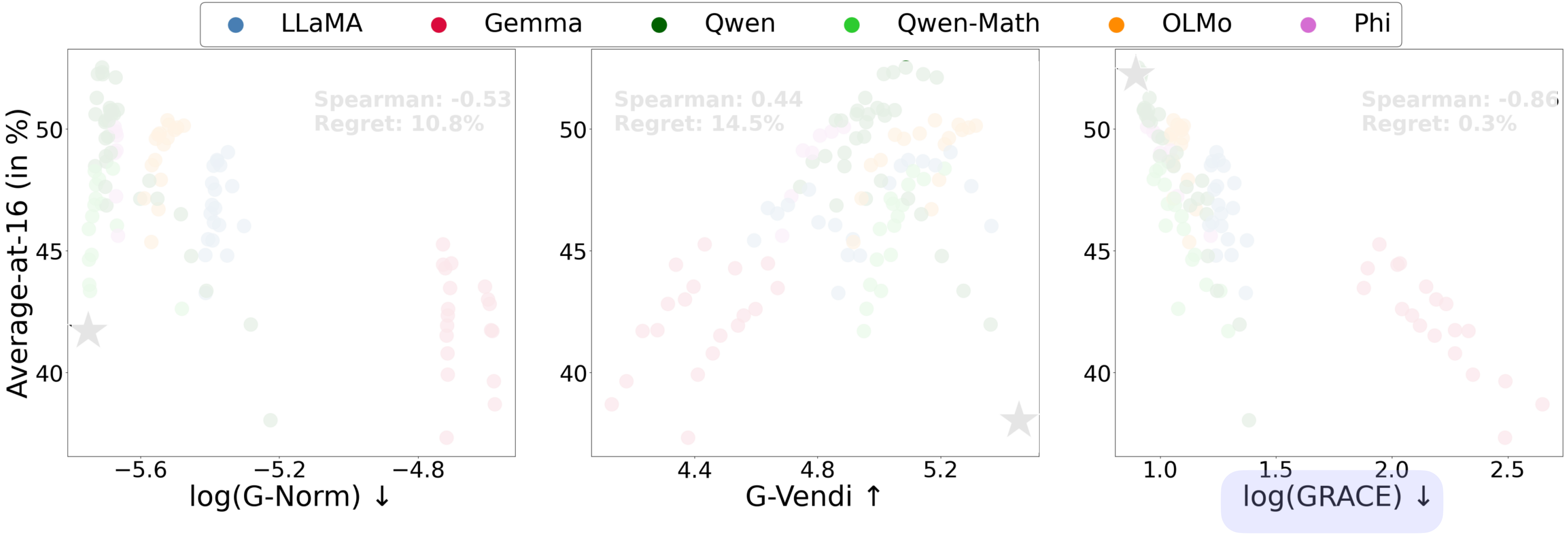
$$\mathbf{GRACE}(D) := \hat{\mathbb{E}}_{D_1 \cup D_2 = D}\left[\mathrm{Tr}\left(\Sigma_1 \tilde{\Sigma}_2^{-1}\right)\right]$$

1) **Indicative of the student's performance.**

2) Informing distillation practices.

GRACE: high correlation, low regret.

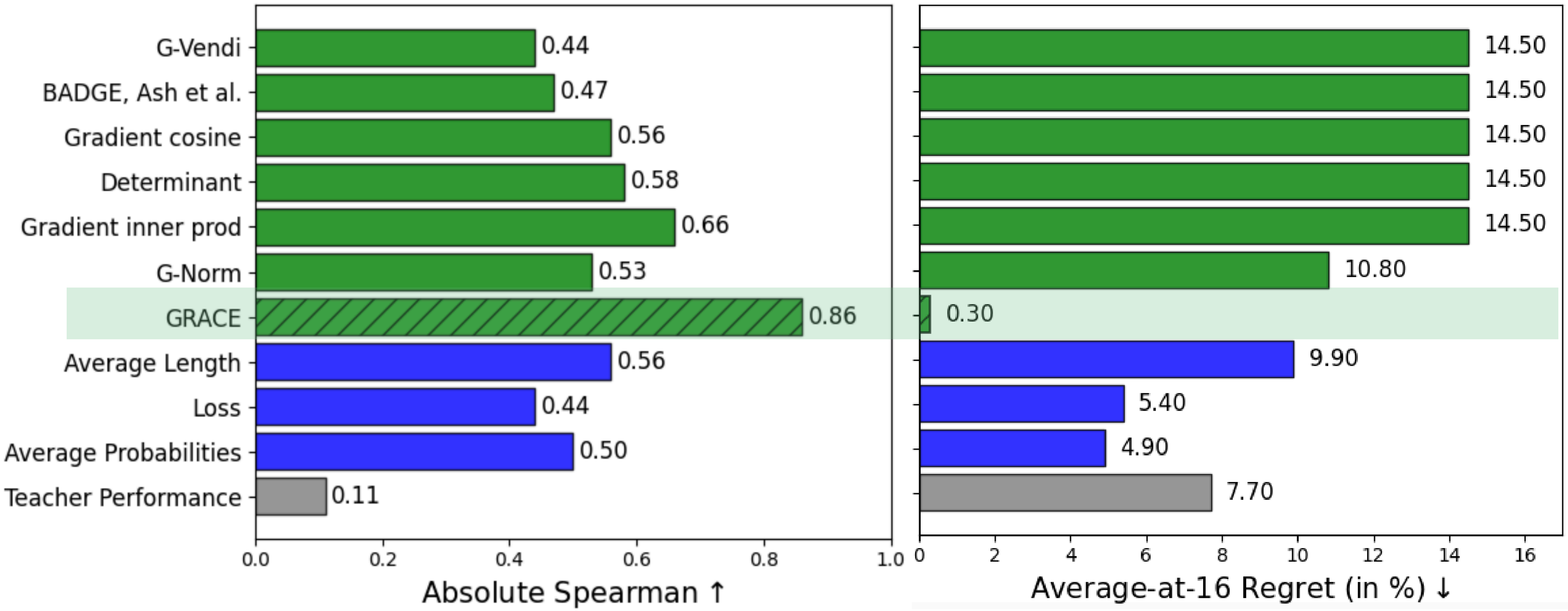
LLaMA-1B on GSM8K *Similar results for other students and datasets (MATH and OOD).



GRACE: high correlation, low regret

LLaMA-1B on GSM8K

*Similar results for other students and datasets (MATH and OOD).



GRACE for LLM teacher selection

(GRAdient Cross-validation Evaluation)

$$\mathbf{GRACE}(D) := \hat{\mathbb{E}}_{D_1 \cup D_2 = D}\left[\mathrm{Tr}\left(\Sigma_1 \tilde{\Sigma}_2^{-1}\right)\right]$$

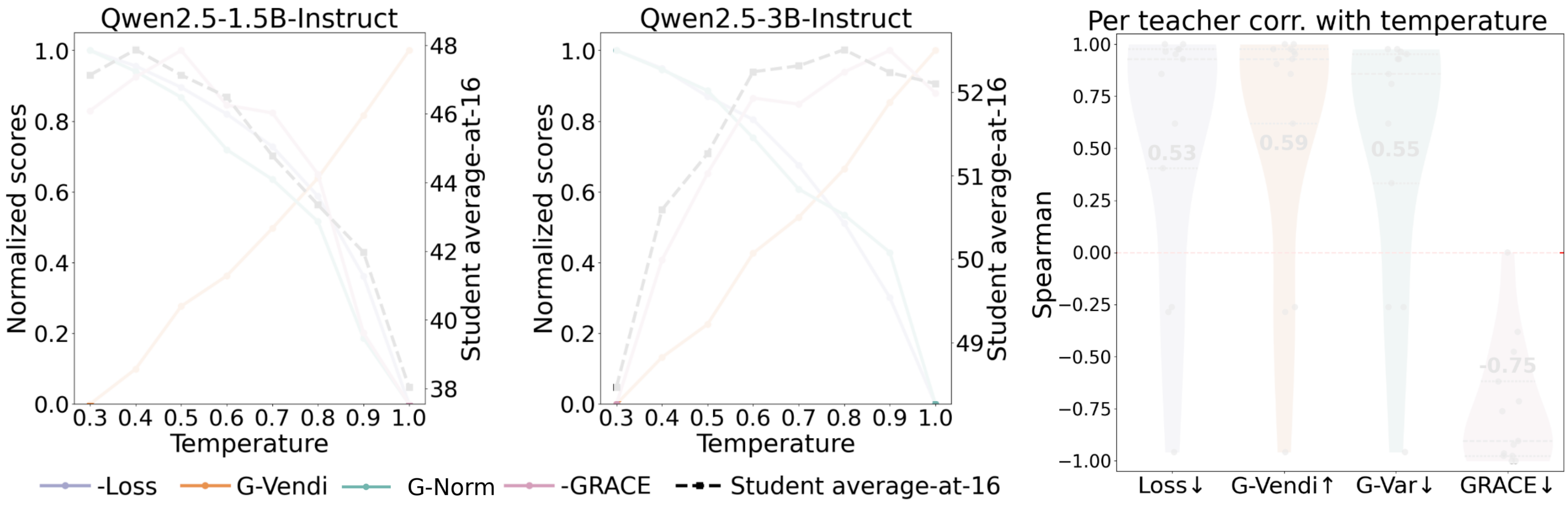
1) Indicative of the student's performance.

2) **Informing distillation practices.**

temperature / size / family

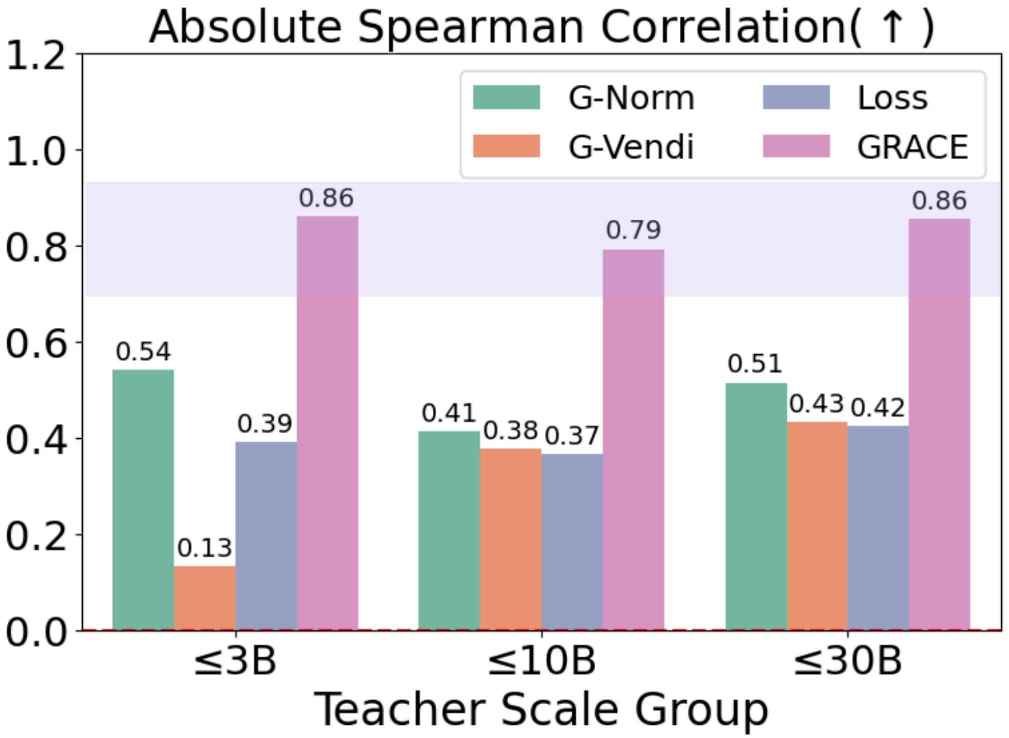
GRACE guides distillation practice

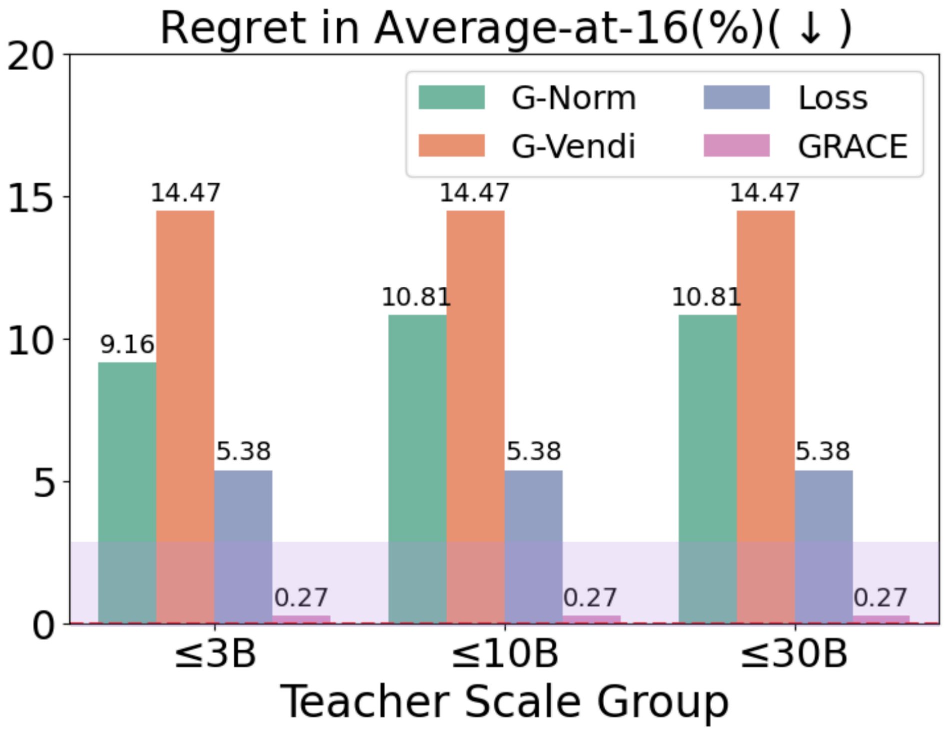
1. Selecting teacher generation temperature.



GRACE guides distillation practice

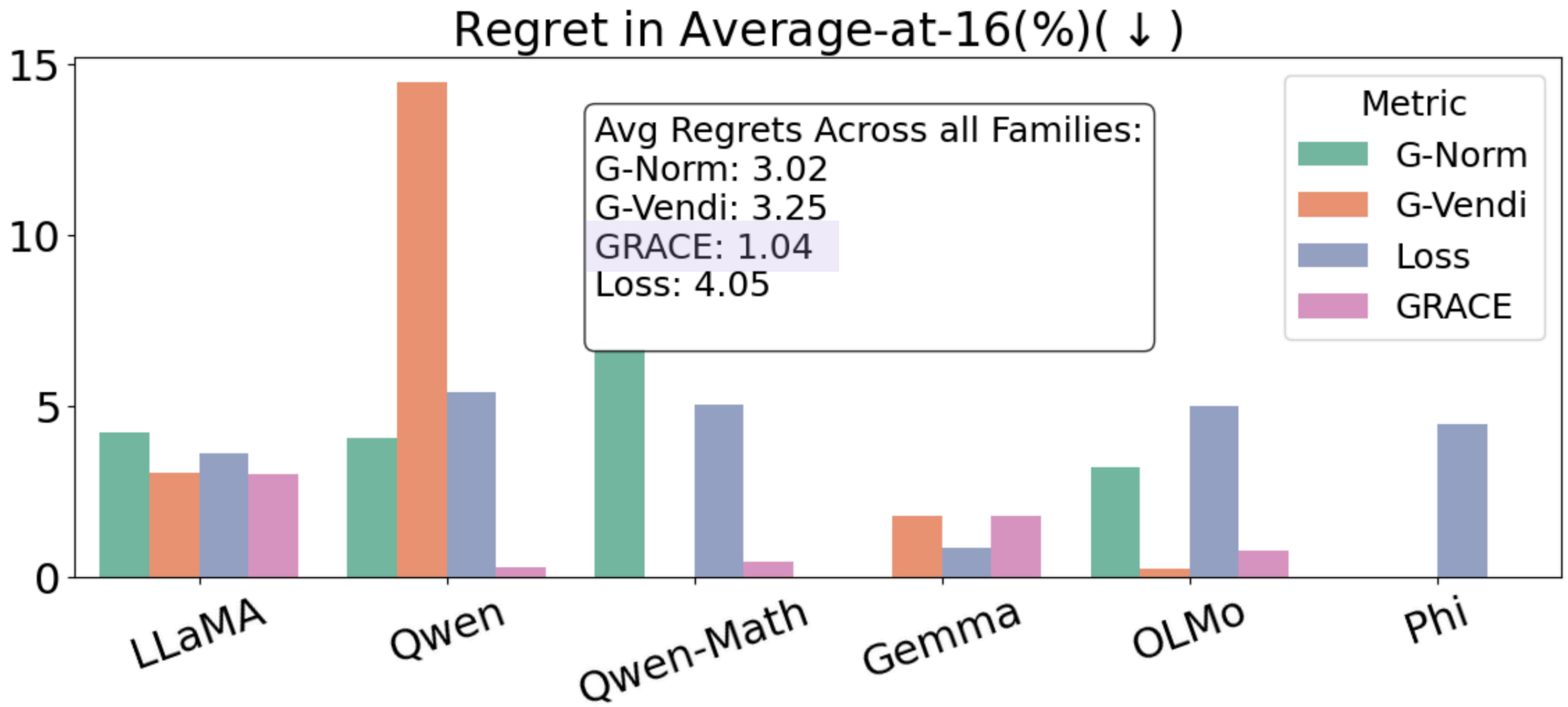
2. Selecting teacher under size constraints.





GRACE guides distillation practice

3. Selecting within a model family.



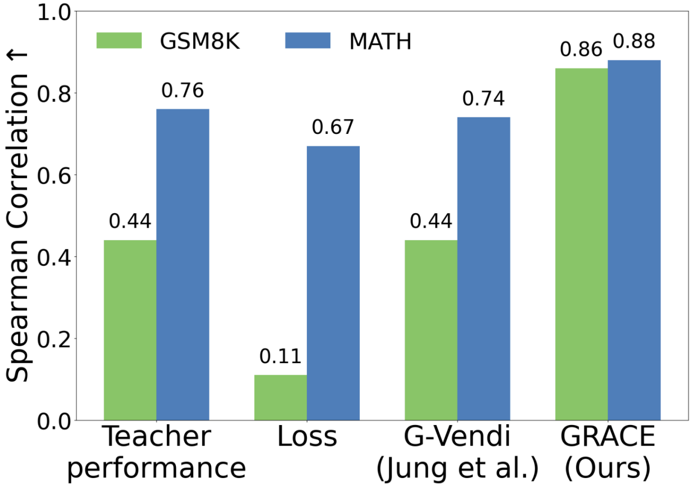
Part 2: teacher selection for LLM post-training

GRACE (~spectrum-weighted norm):

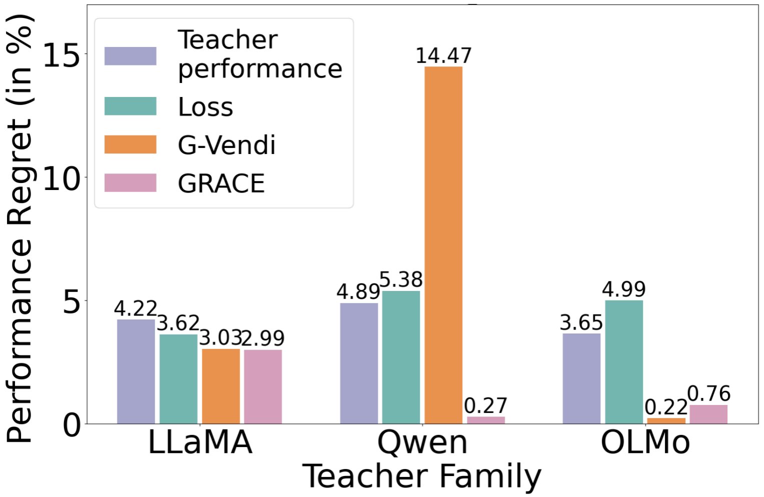
1) Reliably selecting a good teacher;

2) Informing distillation practices.
(temperature / size / family)

high correlation



low regret



Distilling from the *right* teacher

Progressive distillation: implicit curricula from *intermediate checkpoints*.

- Case study on sparse parity: improved sample complexity.
- Empirically verified on PCFG and language experiments.

GRACE for teacher selection: scoring teachers for LLM post-training.

- Indicative of student's performance … high correlation, low regret.
- Guide design choices … temperature, under size-constraints, within a family.

Takeaway: a lot to gain from distillation

Proper design choices matter: **right teacher**, algorithm, understanding.

Many interesting questions & connections.

- Learning with dense supervision / CoT [Joshi et al. 25, Kim et al. 25]
- RL: imitation learning / behavior cloning [Rohatgi et al. 25]
- Model stealing [Liu & Moitra 24]

Practical impact.

- Variants, e.g. on-policy distillation [Qwen 3, Thinking Machine blog].
- Weight distillation as better initialization [Bick et al. 24, Wang et al. 24].

Appendix

Thank you for wanting to know more :)

Better teacher $\nrightarrow$ better student

"capacity gap"
(due to differences in size / training steps)

Model	Dataset	BLKD	TAKD
CNN	CIFAR-10	72.57	73.51
	CIFAR-100	44.57	44.92
ResNet	CIFAR-10	88.65	88.98
	CIFAR-100	61.41	61.82
ResNet	ImageNet	66.60	67.36

[Mirzadeh et al. 19]

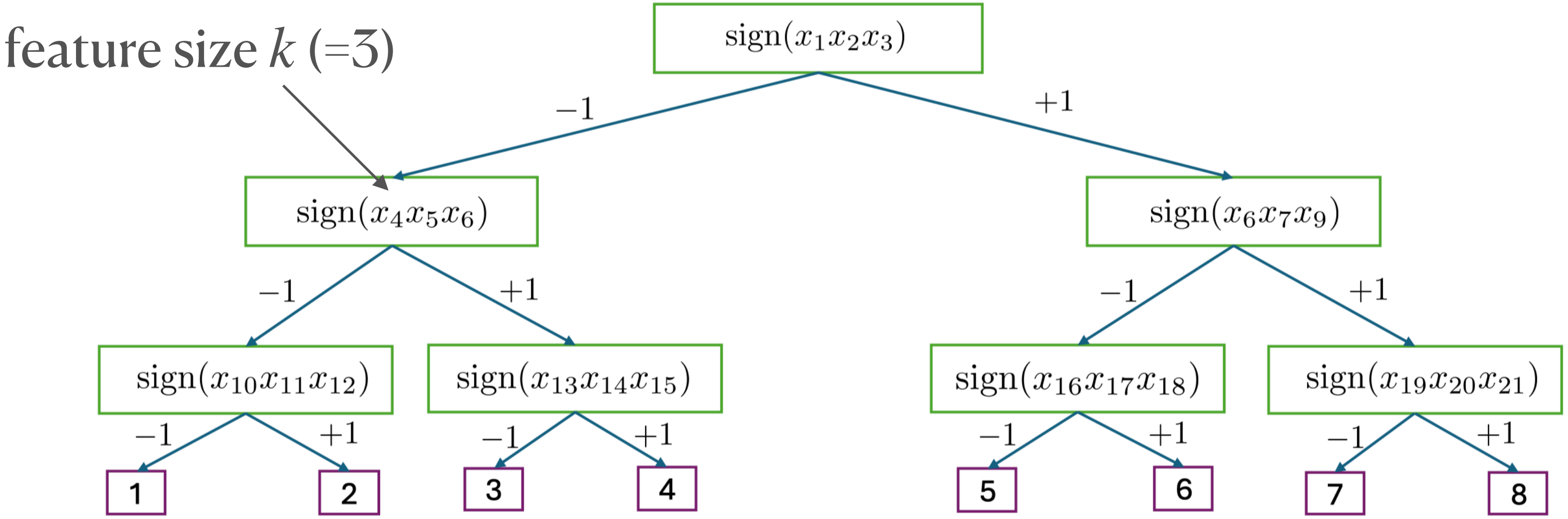
CIFAR-100

ResNet-56 $\rightarrow$ LeNet-5x8	50.1 ± 0.4	61.9 ± 0.2
ResNet-56 $\rightarrow$ ResNet-20	68.2 ± 0.3	69.6 ± 0.3
ResNet-110 $\rightarrow$ LeNet-5x8	48.6 ± 0.8	60.8 ± 0.2
ResNet-110 $\rightarrow$ ResNet-20	67.8 ± 0.2	69.0 ± 0.3

[Harutyunyan et al. 23]

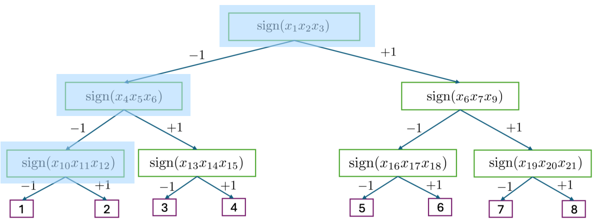
Beyond sparse parity — a hierarchical task

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

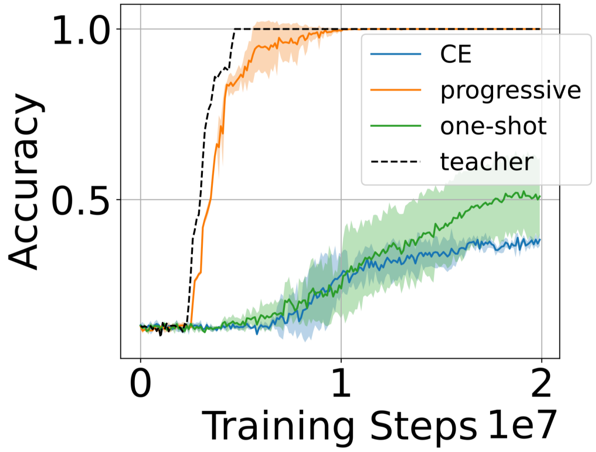


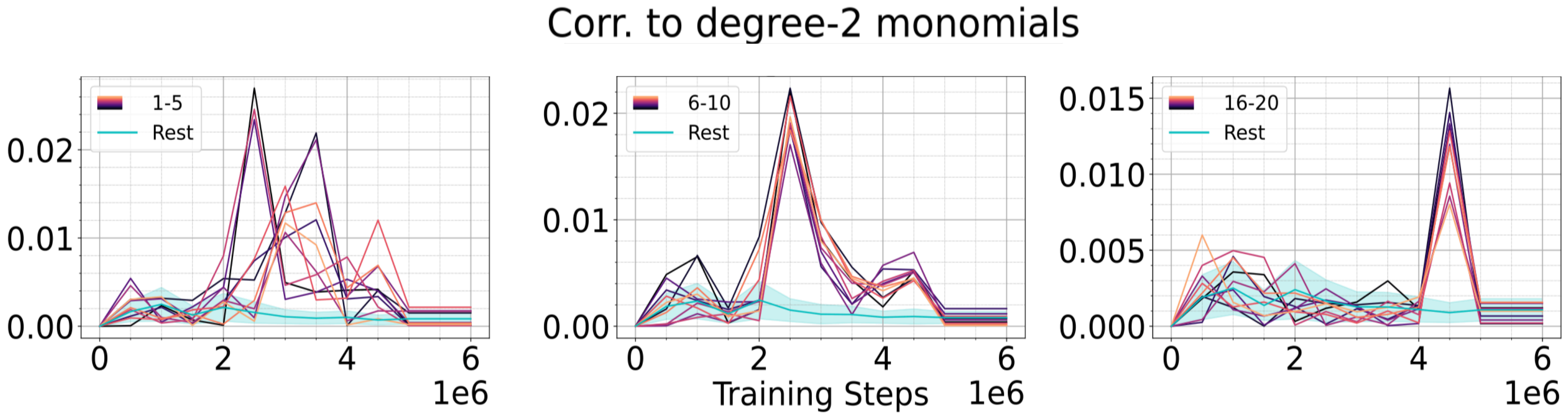
Beyond sparse parity — a hierarchical task

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



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

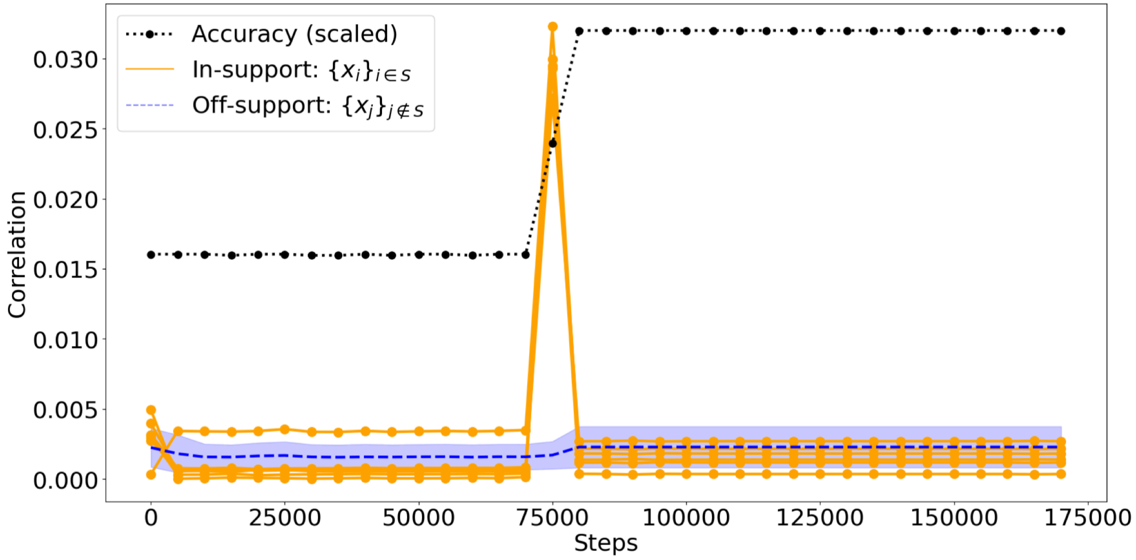


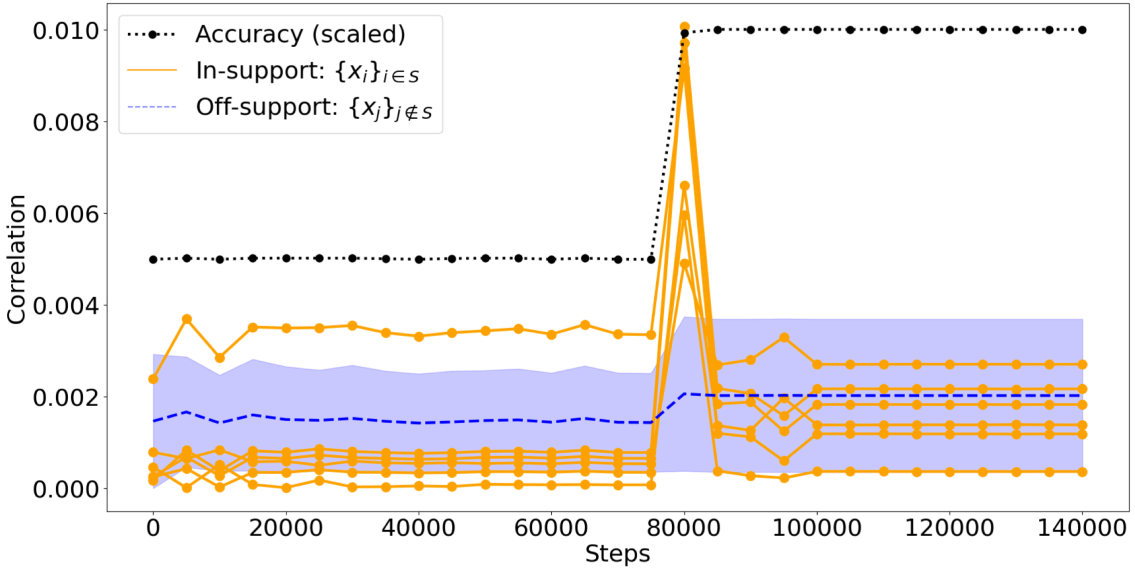


Learning at diff speed → need multiple teachers.

Transformer on sparse parity

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

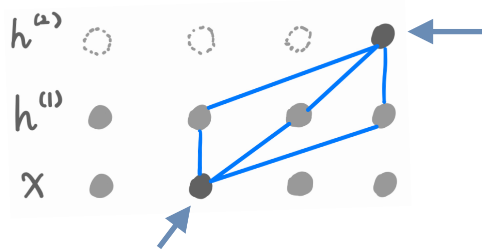


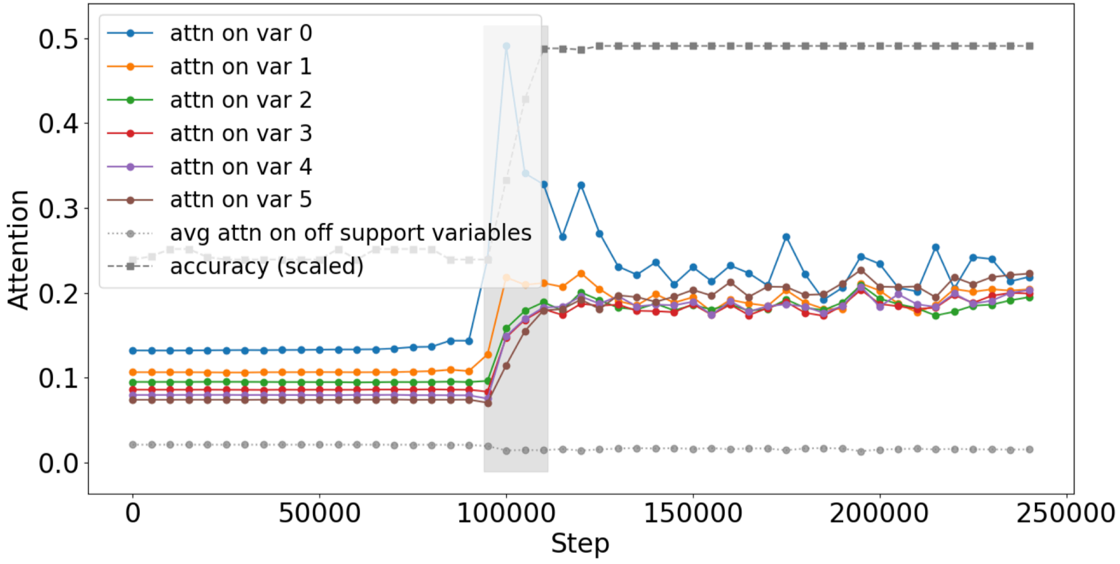


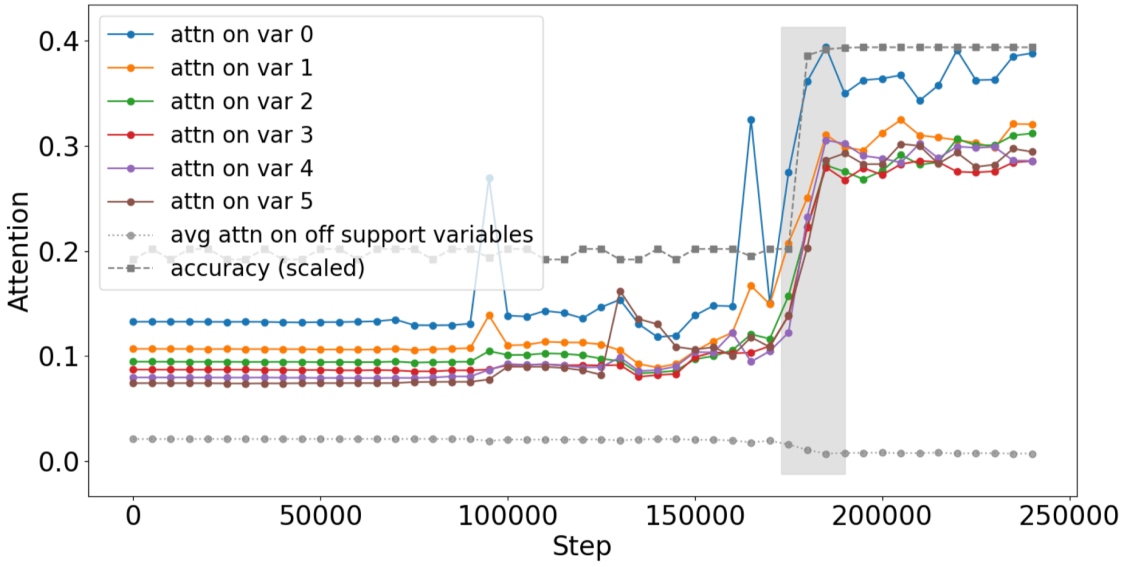
Transformer on sparse parity

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

sum of length-2 paths



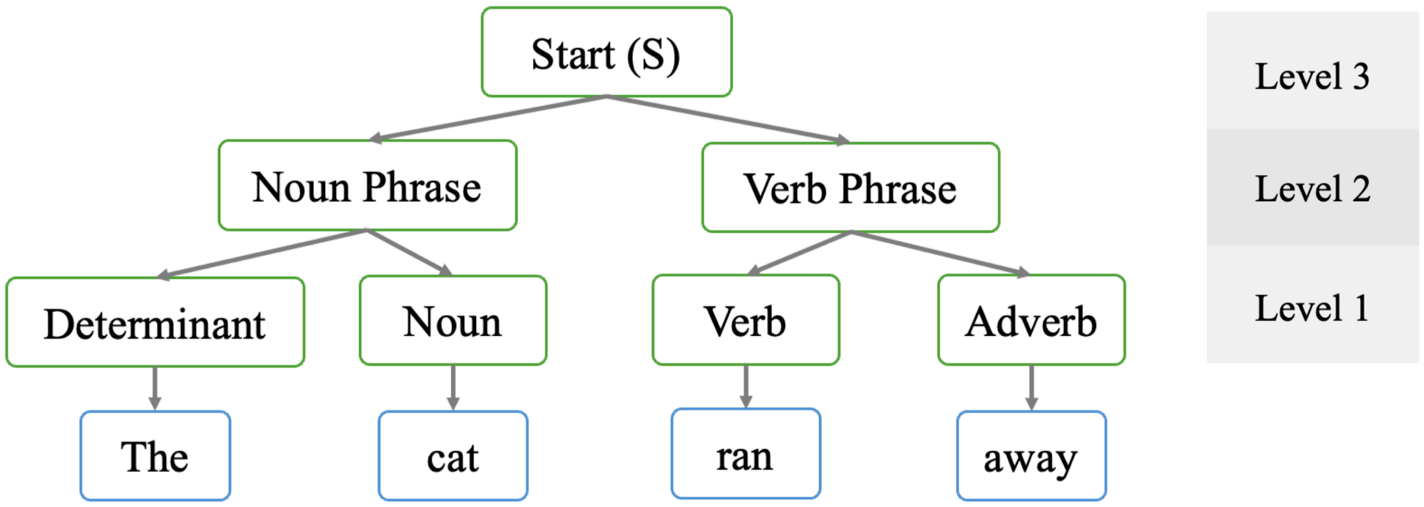




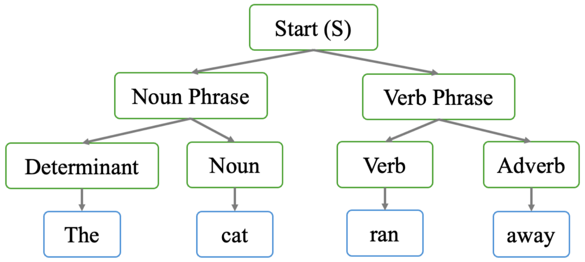
Beyond sparse parity — PCFG

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

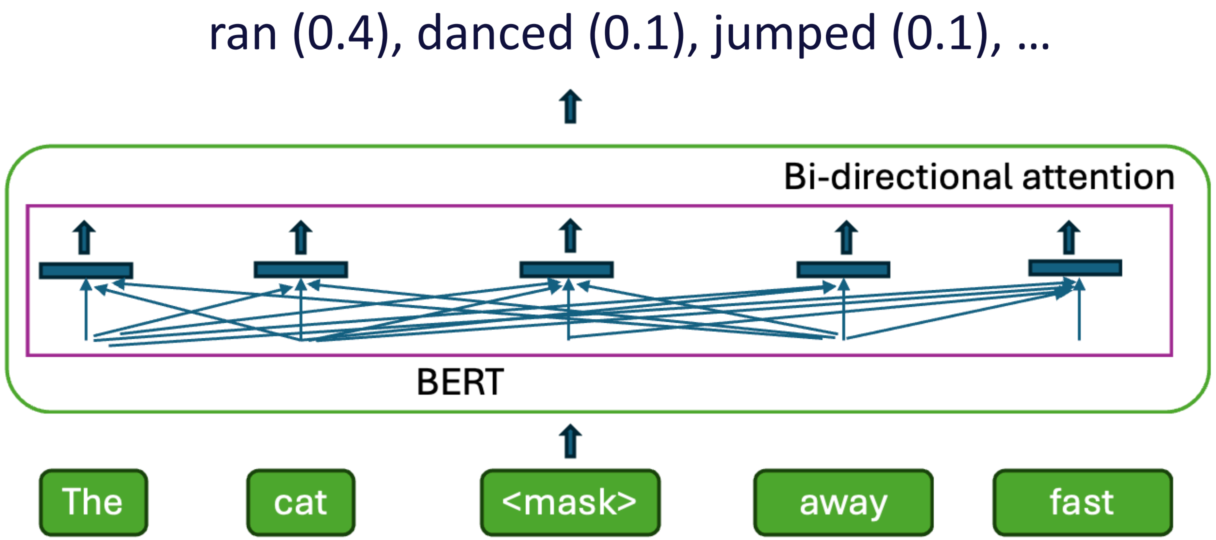
- Defined by: vocab; non-terminals; rules & probabilities.



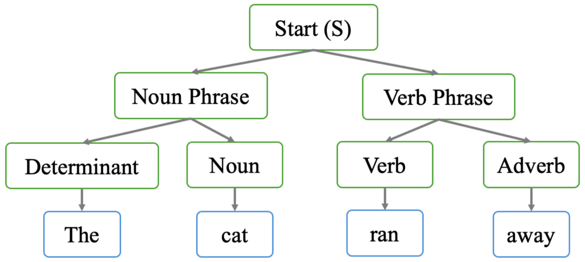
Beyond sparse parity — PCFG



Task: masked prediction ... loss averaged over the masked set.



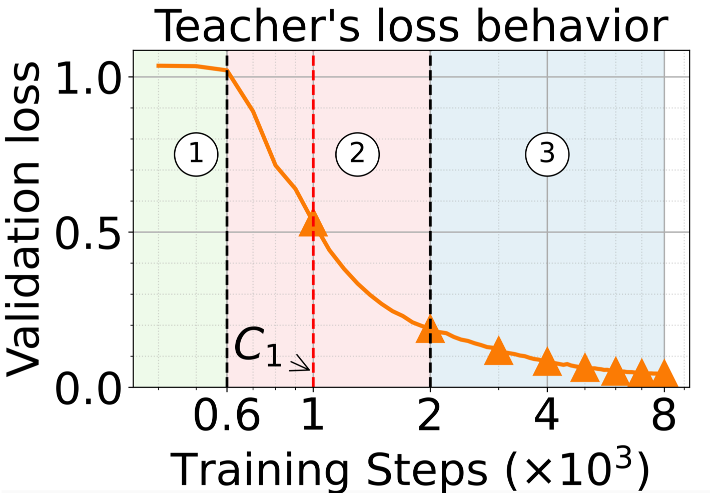
Beyond sparse parity — PCFG

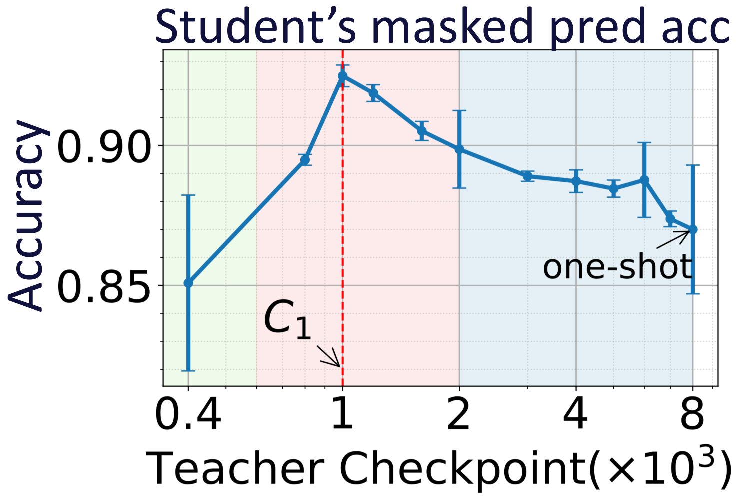


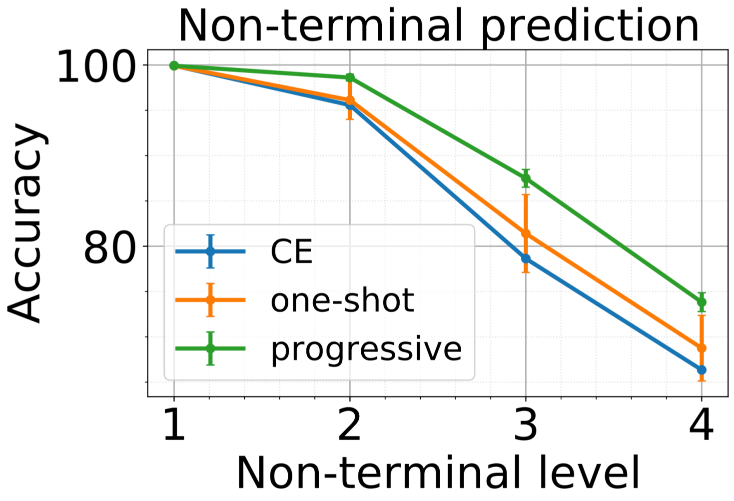
Task: masked prediction → optimal: following the tree hierarchy [Zhao et al. 23].

a quality measure

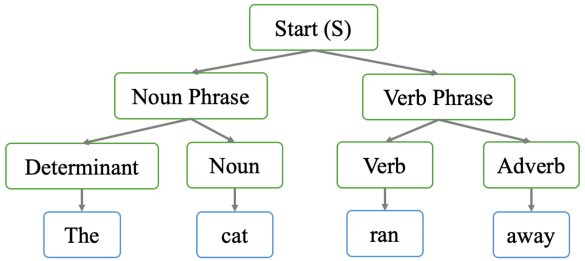
An implicit curriculum exists. *... what is it?*







Implicit curriculum for PCFG



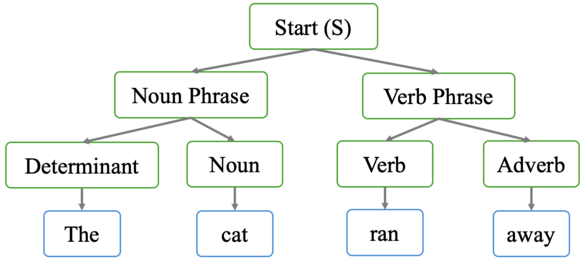
n-grams with an increasing *n*. (e.g. $n = 3$: cat ran away, cat danced away, cat jumped away, …)

- 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:

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

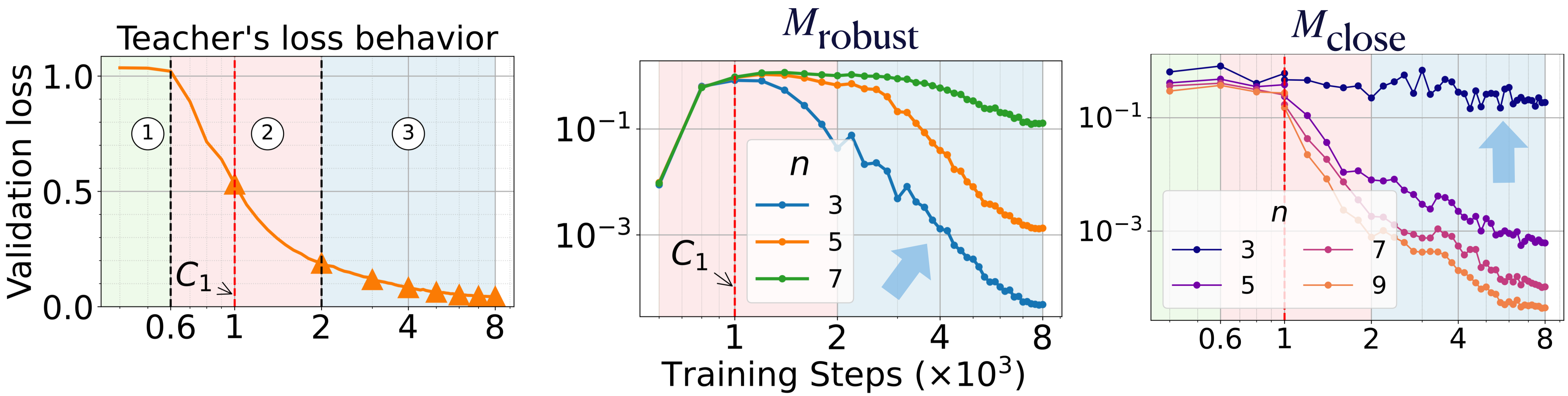
Implicit curriculum for PCFG



n-grams with an increasing *n*.

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

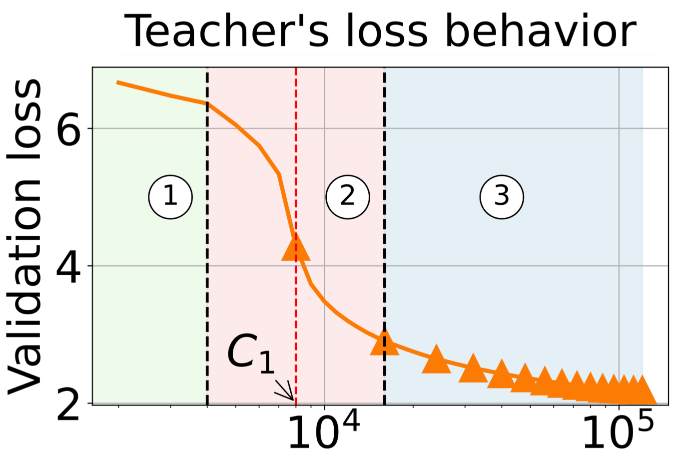
2 measures: later checkpoints depend more on higher *n* (i.e. harder).

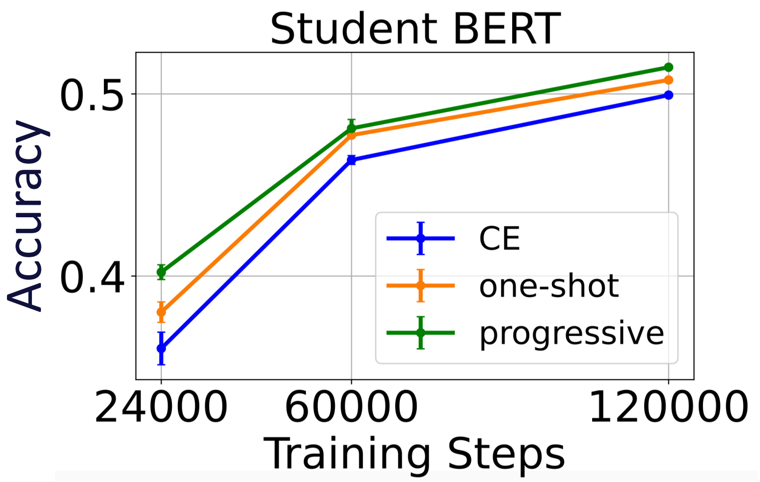


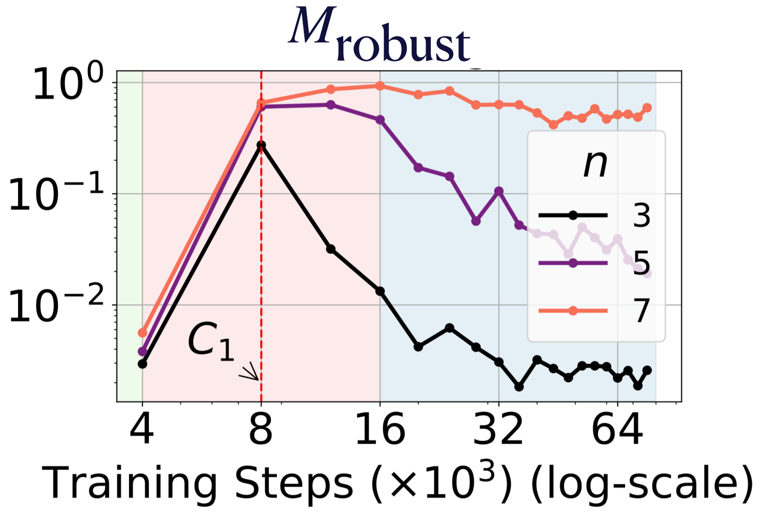
Natural languages

Masked prediction on Wikipedia and Books.

- Similar results for next-token prediction.







Designing GRACE

GRACE: gradient norm, weighted by the spectrum of normalized gradients.

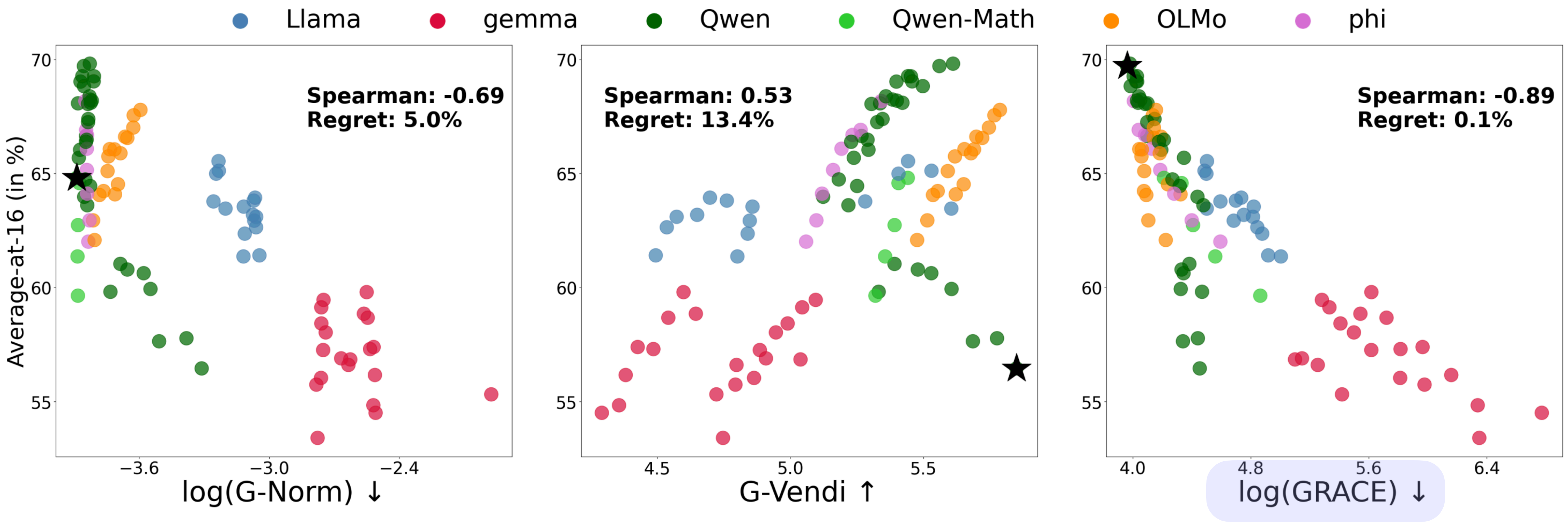
$$\textbf{GRACE}(D) := \hat{\mathbb{E}}_{D_1 \cup D_2 = D}\left[\mathrm{Tr}\left(\Sigma_1 \tilde{\Sigma}_2^{-1}\right)\right]$$

Example:

- Random generations: high G-Norm ↓ and GRACE ↓ (but high G-Vendi ↑).
- Always the same generation: high G-Norm ↓ and GRACE ↓.
- Extreme (random) repetitions: low G-Norm ↓ and GRACE ↓.

GRACE indicative of math performance

Gemma-2B on GSM8K: high correlation ↑, low regret ↓.



GRACE indicative of math performance

Llama-3B on MATH: high correlation ↑, low regret ↓.