Weak-to-Strong Generalization

Behrad Moniri
University of Pennsylvania

STAT 9911 - Principles of AI: LLMs
April 2025

WEAK-TO-STRONG GENERALIZATION: ELICITING STRONG CAPABILITIES WITH WEAK SUPERVISION

Collin Burns* **Pavel Izmailov*** **Jan Hendrik Kirchner*** **Bowen Baker*** **Leo Gao***

Leopold Aschenbrenner* **Yining Chen*** **Adrien Ecoffet*** **Manas Joglekar***

Jan Leike **Ilya Sutskever** **Jeff Wu***

OpenAI

Alignment

- As models become more capable, alignment will become increasingly critical.

- More capable the models are -> higher stakes.

- Out techniques to make models aligned: RLHF, DPO, PPO, etc.

- Humans are teaching the LLMs to be more aligned.

Alignment

- **Right now**: anyone could help align the model.
- **Eventually**: we will need experts to align the model.
- **In the (distant?) future**: there is no expert anymore!

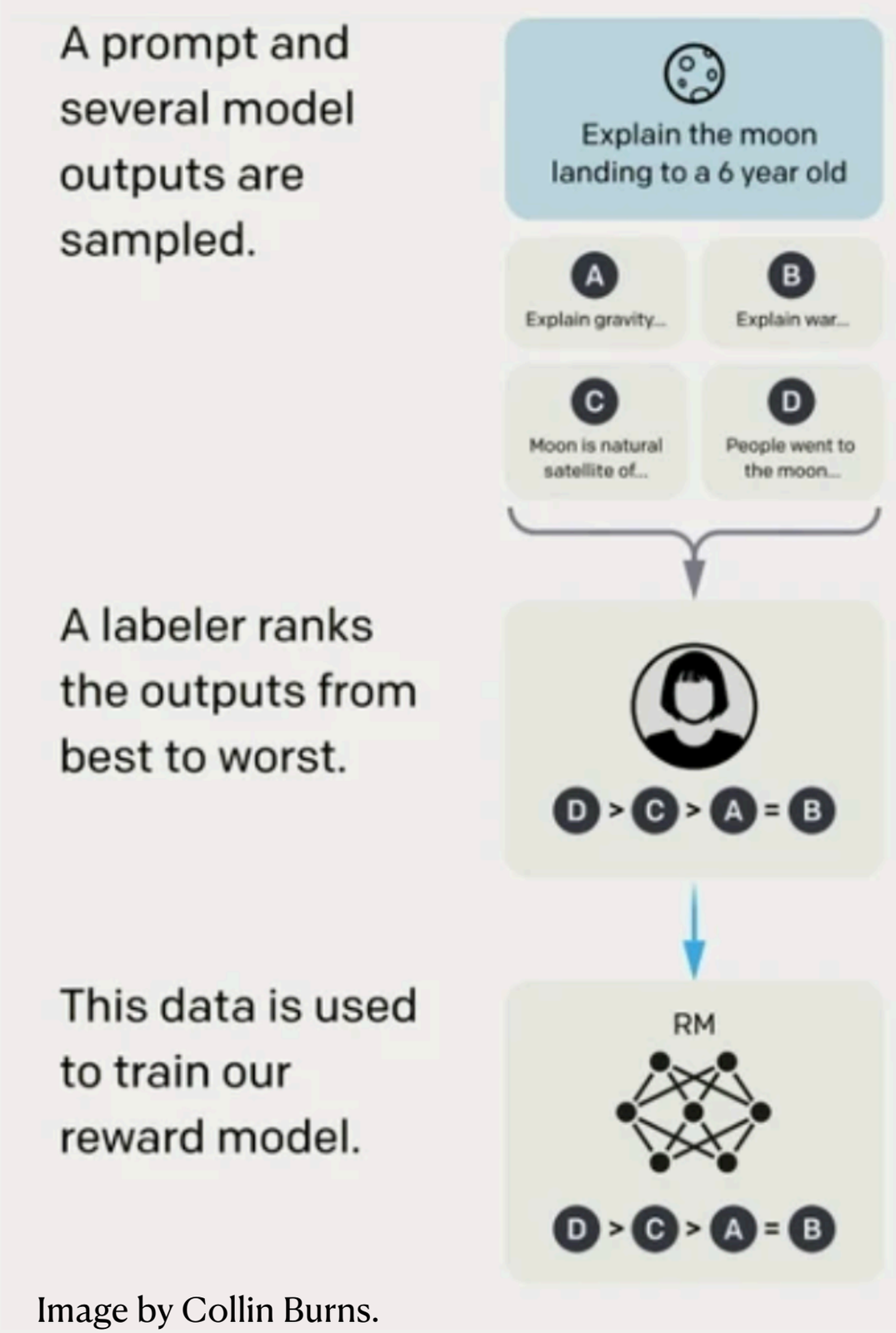


Image by Collin Burns.

SuperAlignment

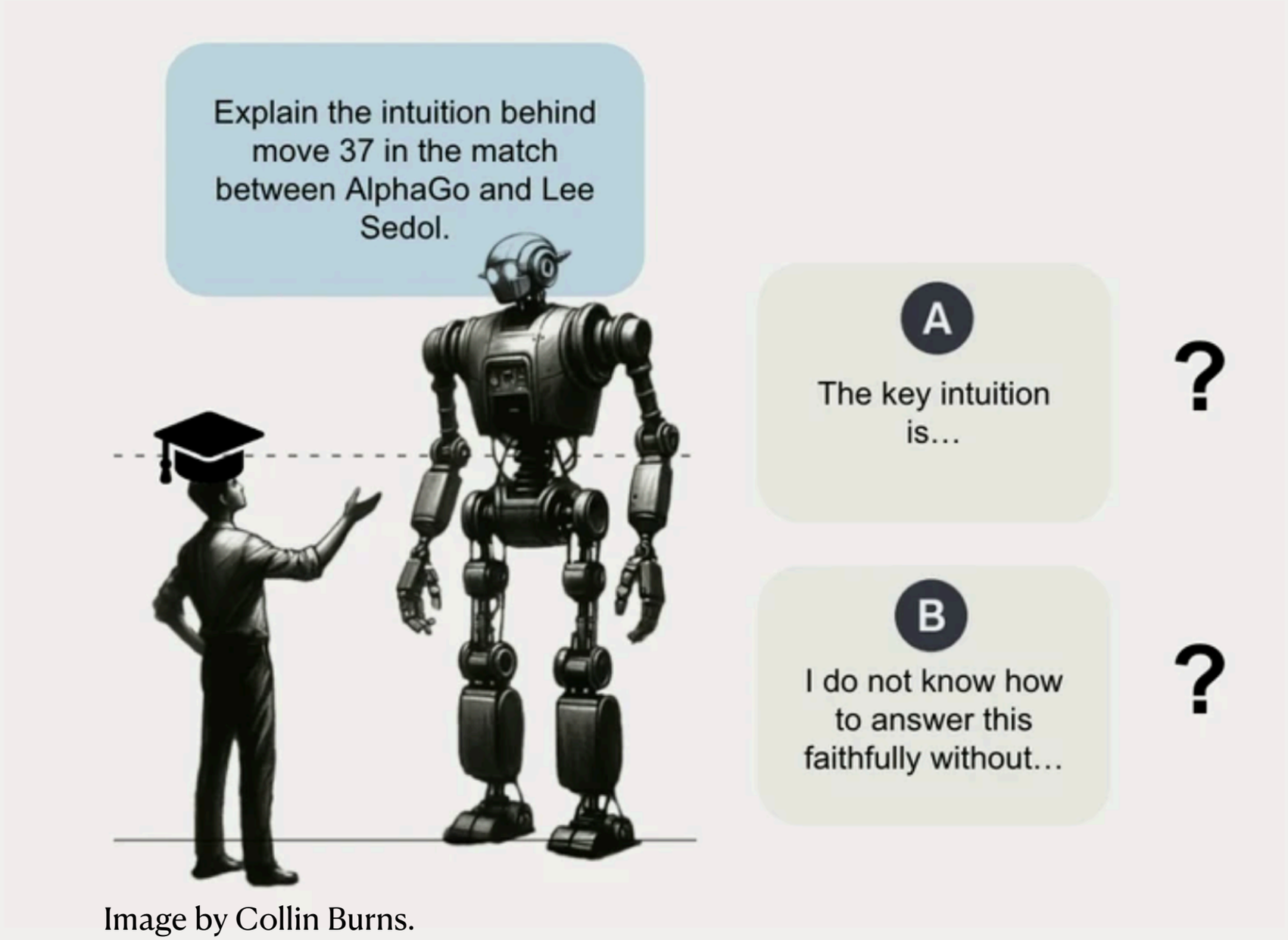


Image by Collin Burns.

The role of humans?

- **Key Question**: Can humans still play a role in training/alignment of the model when the models become super strong?
- This paper aims to answer this question.

- But we don’t have access to superhuman models now! How can we answer this question?

- Weak model vs. Strong Model

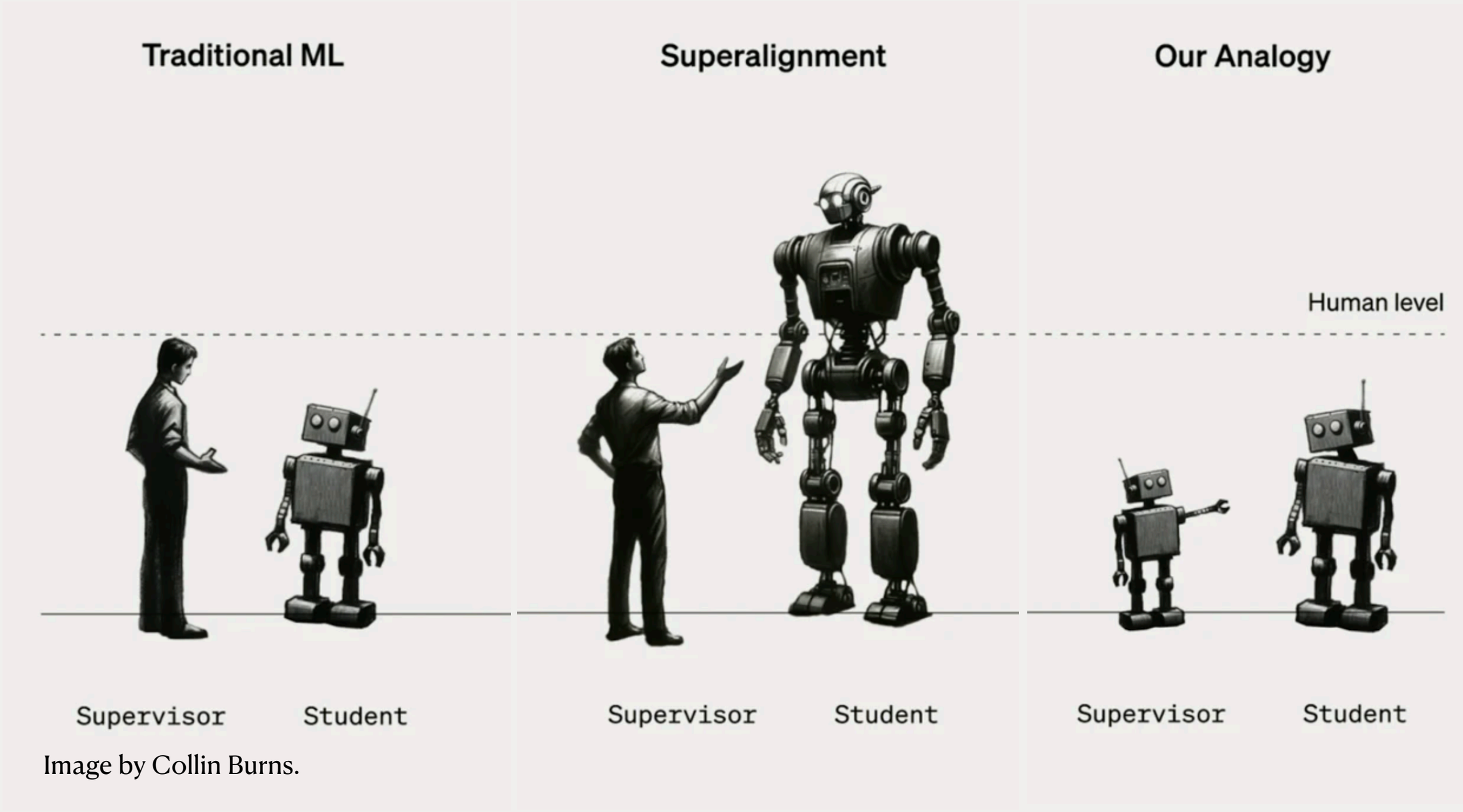


Image by Collin Burns.

Weak-to-Strong Generalization

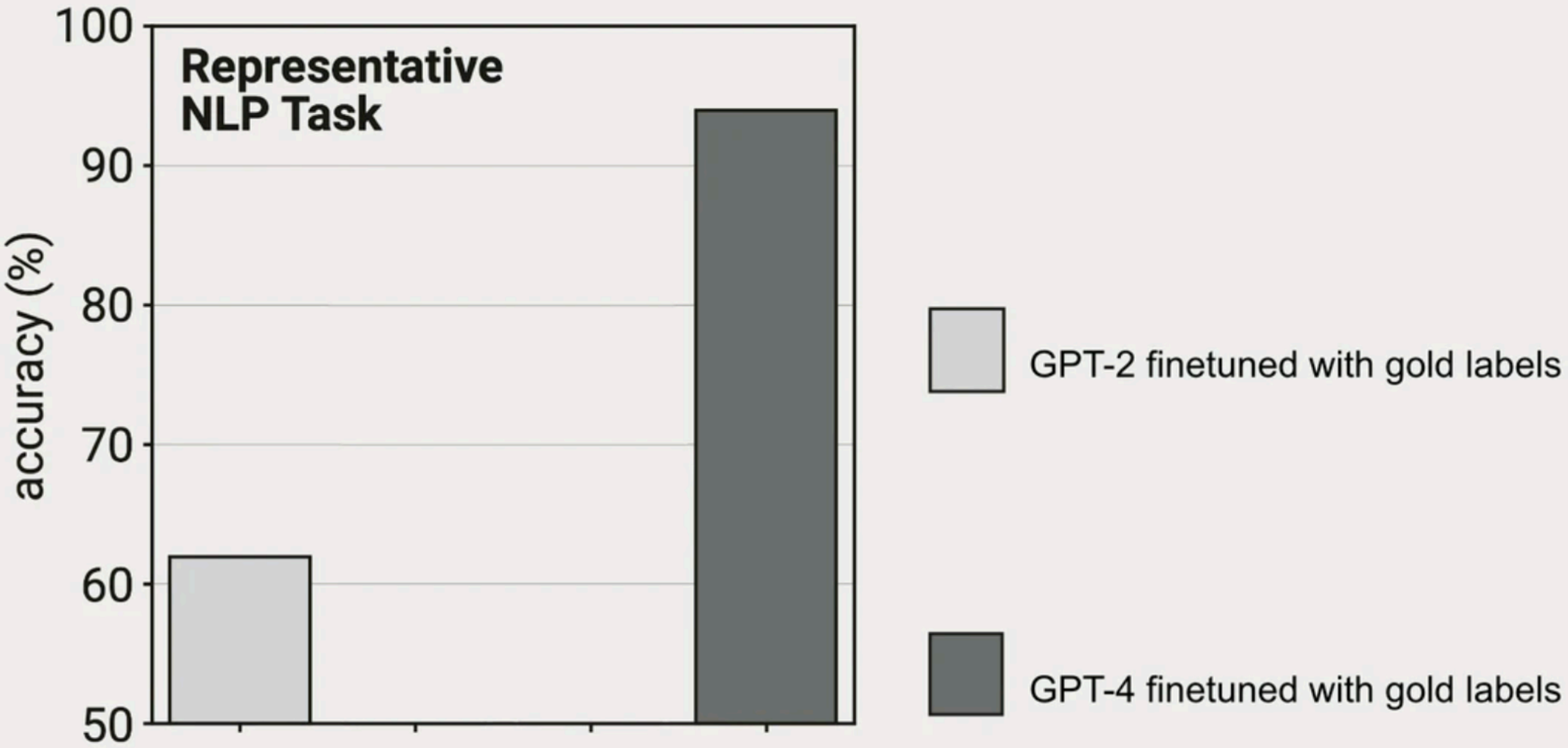


Image by Collin Burns.

Weak-to-Strong Generalization

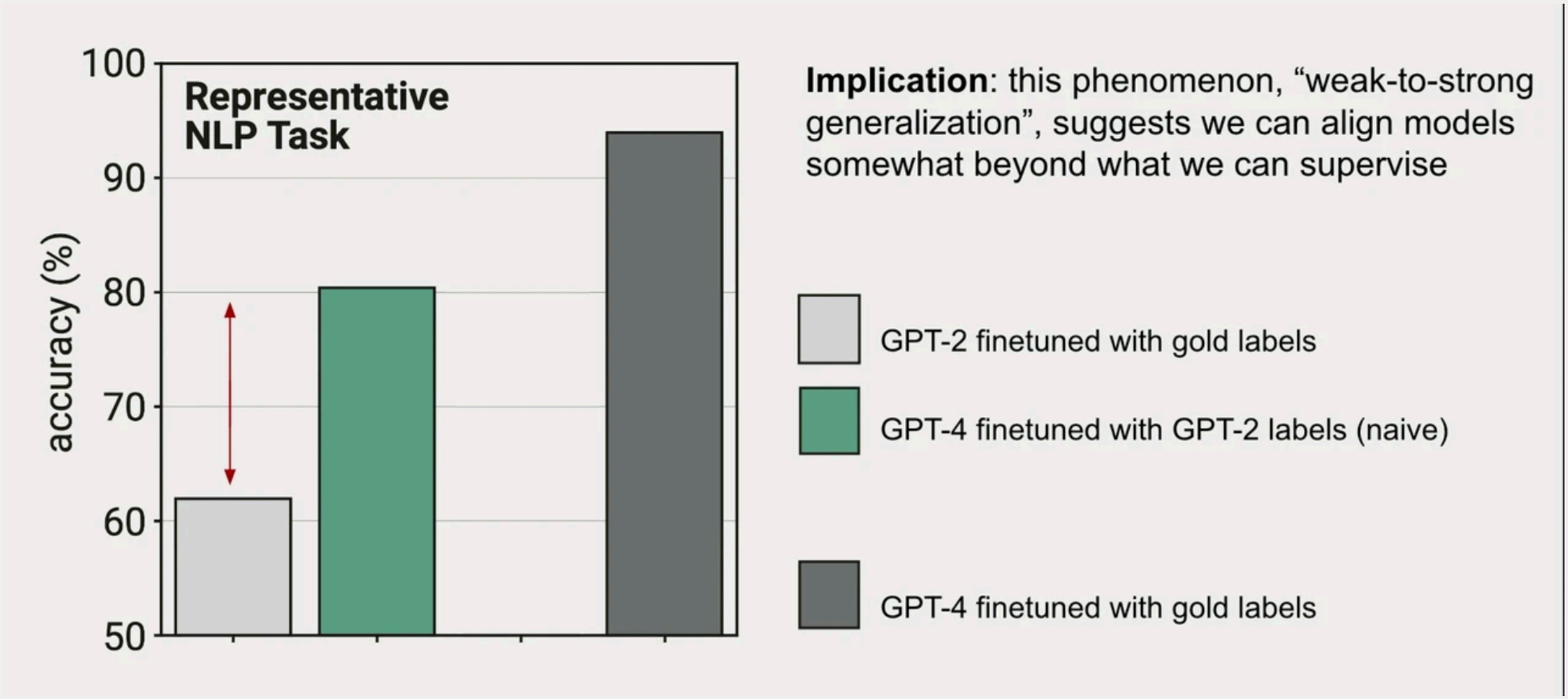
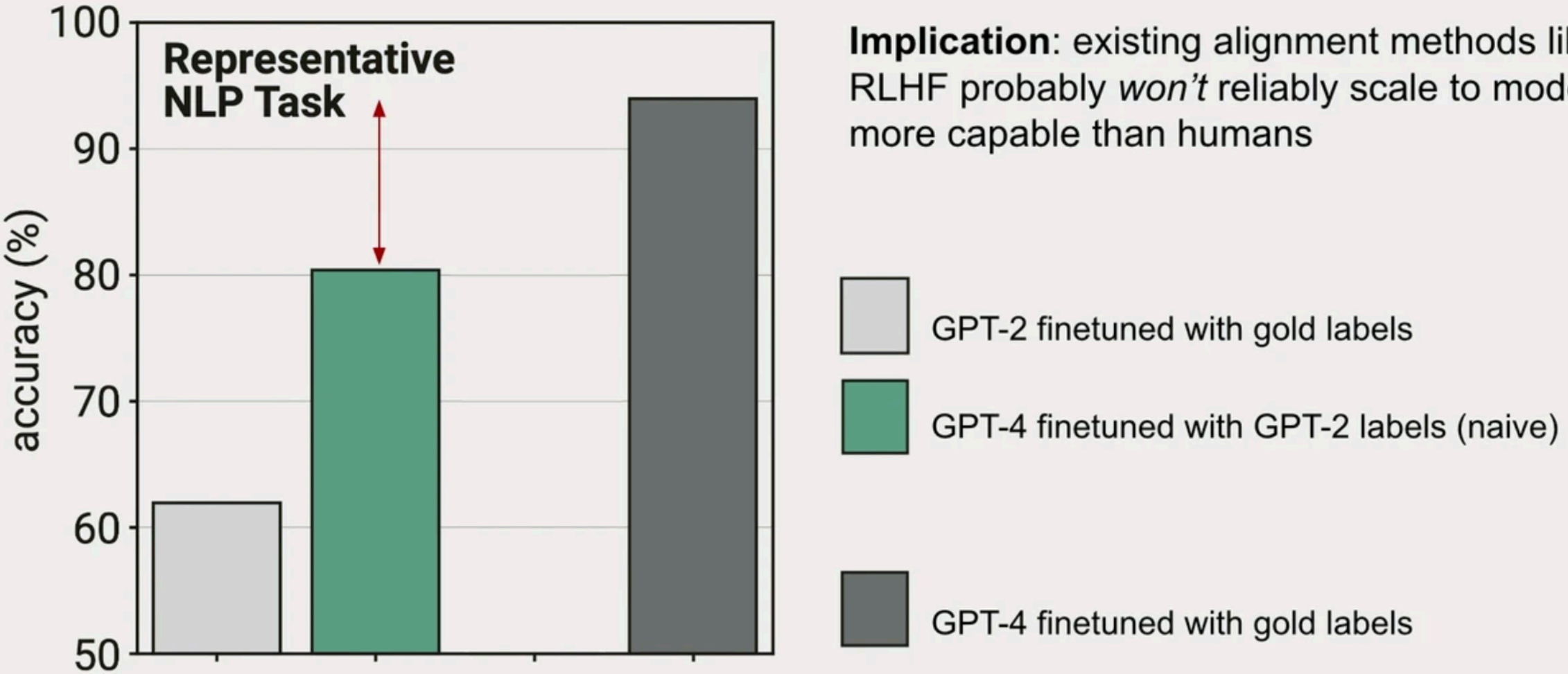


Image by Collin Burns.

Weak-to-Strong Generalization



Implication: existing alignment methods like RLHF probably *won't* reliably scale to models more capable than humans

Image by Collin Burns.

Weak-to-Strong Generalization

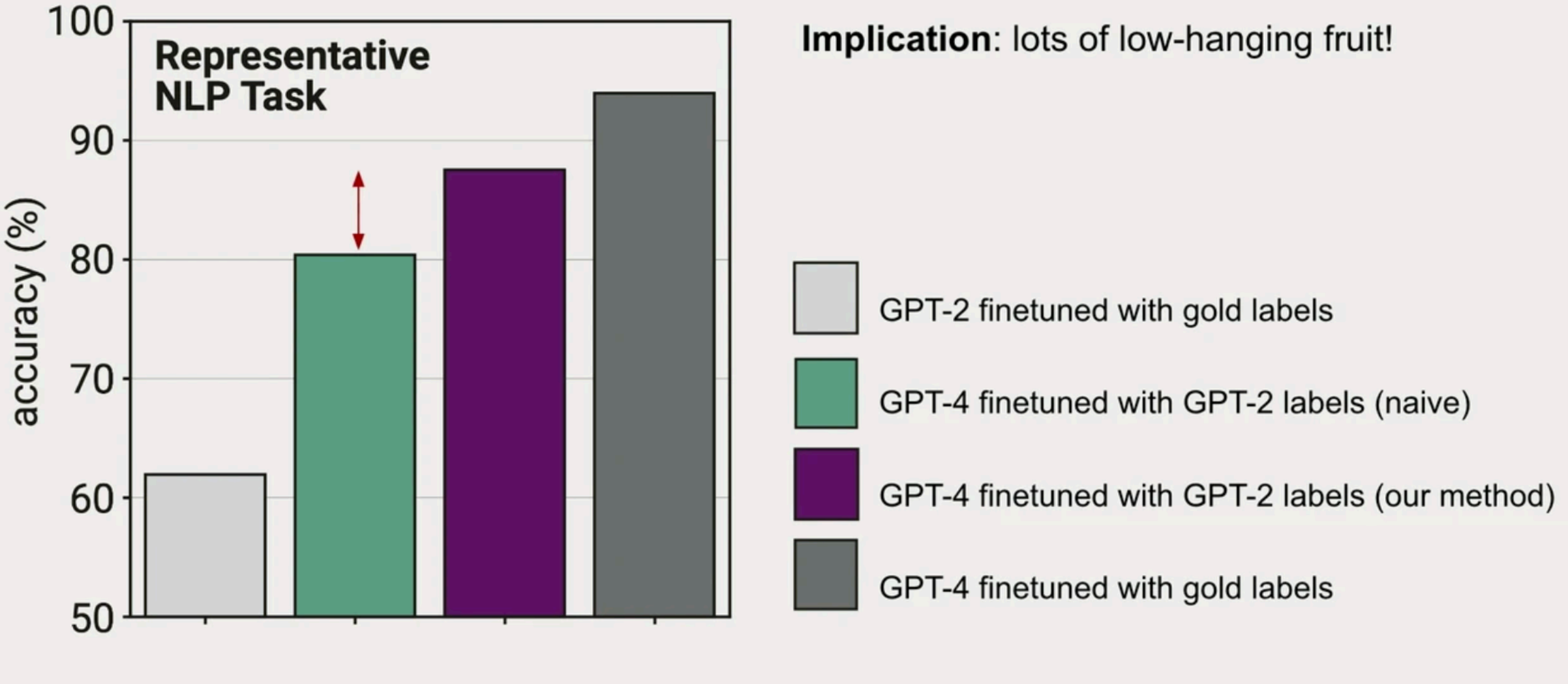


Image by Collin Burns.

Datasets

- NLP **classification** datasets covering ethics, commonsense reasoning, natural language inference, sentiment analysis, and other domains
- **Chess puzzles**: Model has to output a sequence of moves.
- **Reward modeling**: the proprietary dataset used to train ChatGPT reward models.

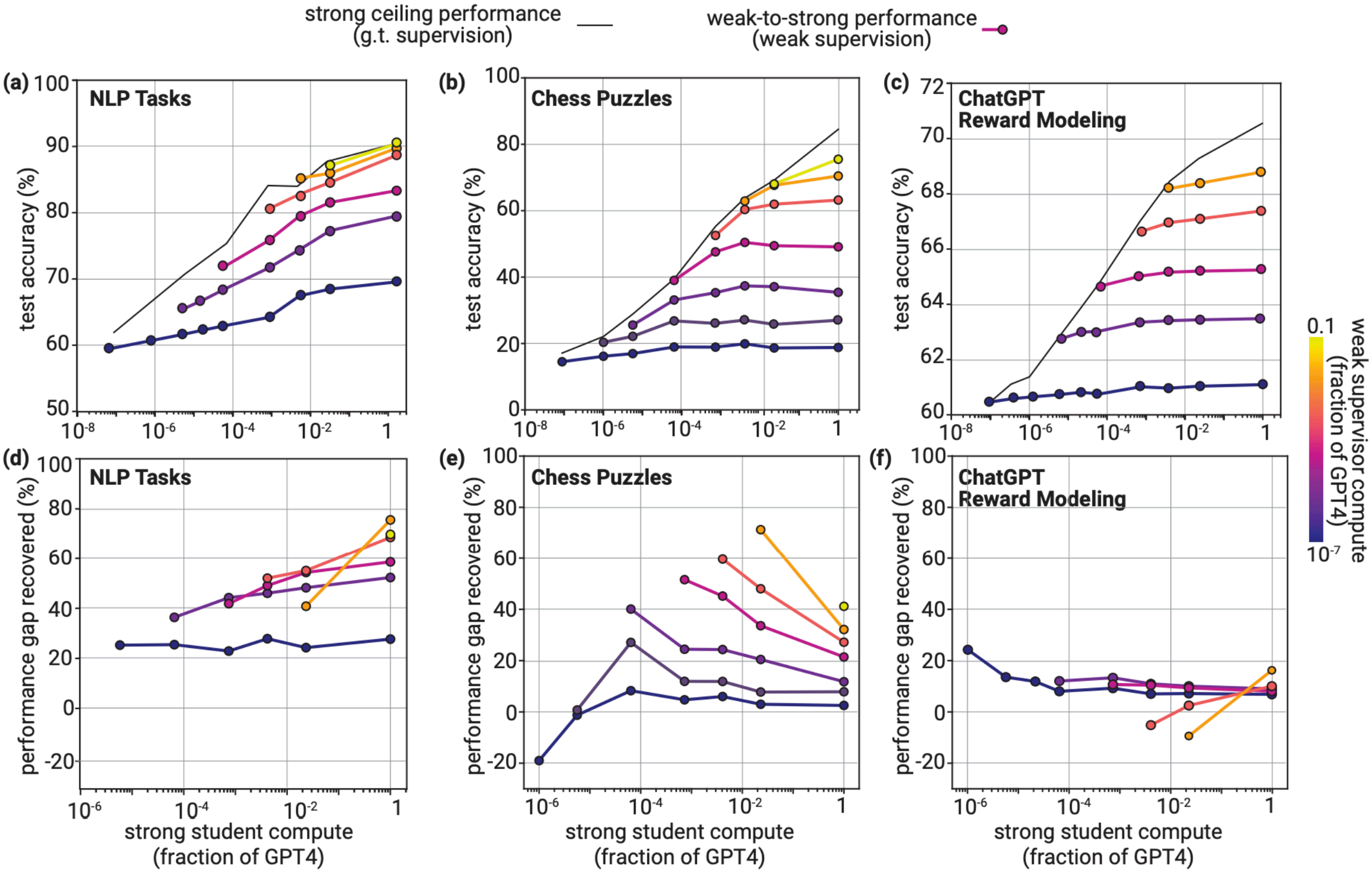
Performance Gap Covered

$$\text{PGR} = \frac{\text{weak-to-strong} - \text{weak}}{\text{strong ceiling} - \text{weak}} = \frac{\text{——}}{\text{——}}$$

weak performance

weak-to-strong performance

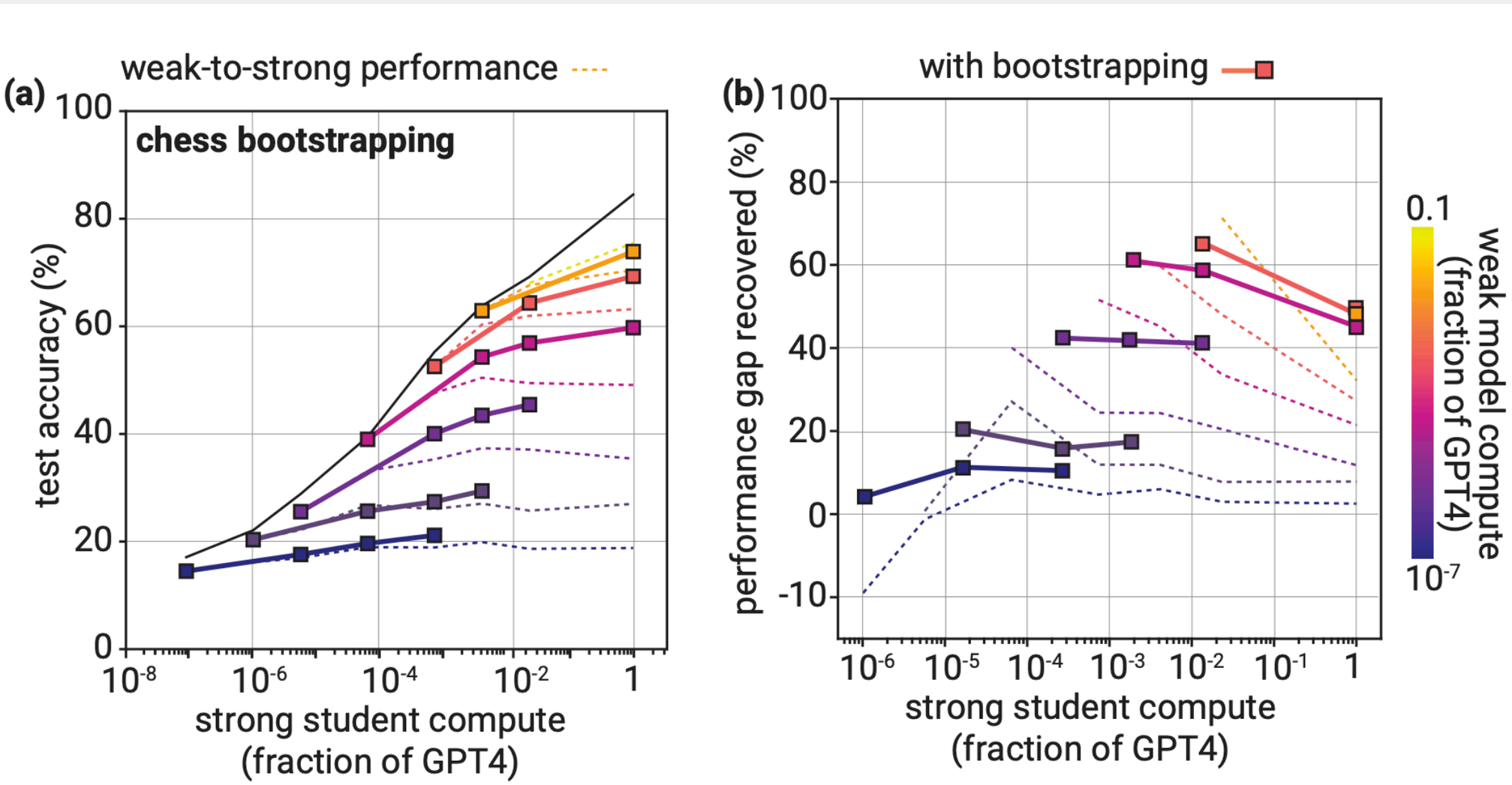
strong ceiling performance


strong ceiling performance
(g.t. supervision)
weak-to-strong performance
(weak supervision)
(a)
NLP Tasks
(b)
Chess Puzzles
(c)
ChatGPT
Reward Modeling
test accuracy (%)
(d)
NLP Tasks
(e)
Chess Puzzles
(f)
ChatGPT
Reward Modeling
performance gap recovered (%)
strong student compute
(fraction of GPT4)
weak supervisor compute
(fraction of GPT4)
0.1
10^{-7}

Improving Weak-to-Strong Generalization

Bootstrapping

- Instead of directly aligning very strong models, we could first align an only slightly strong model, use that to align an even stronger models ...

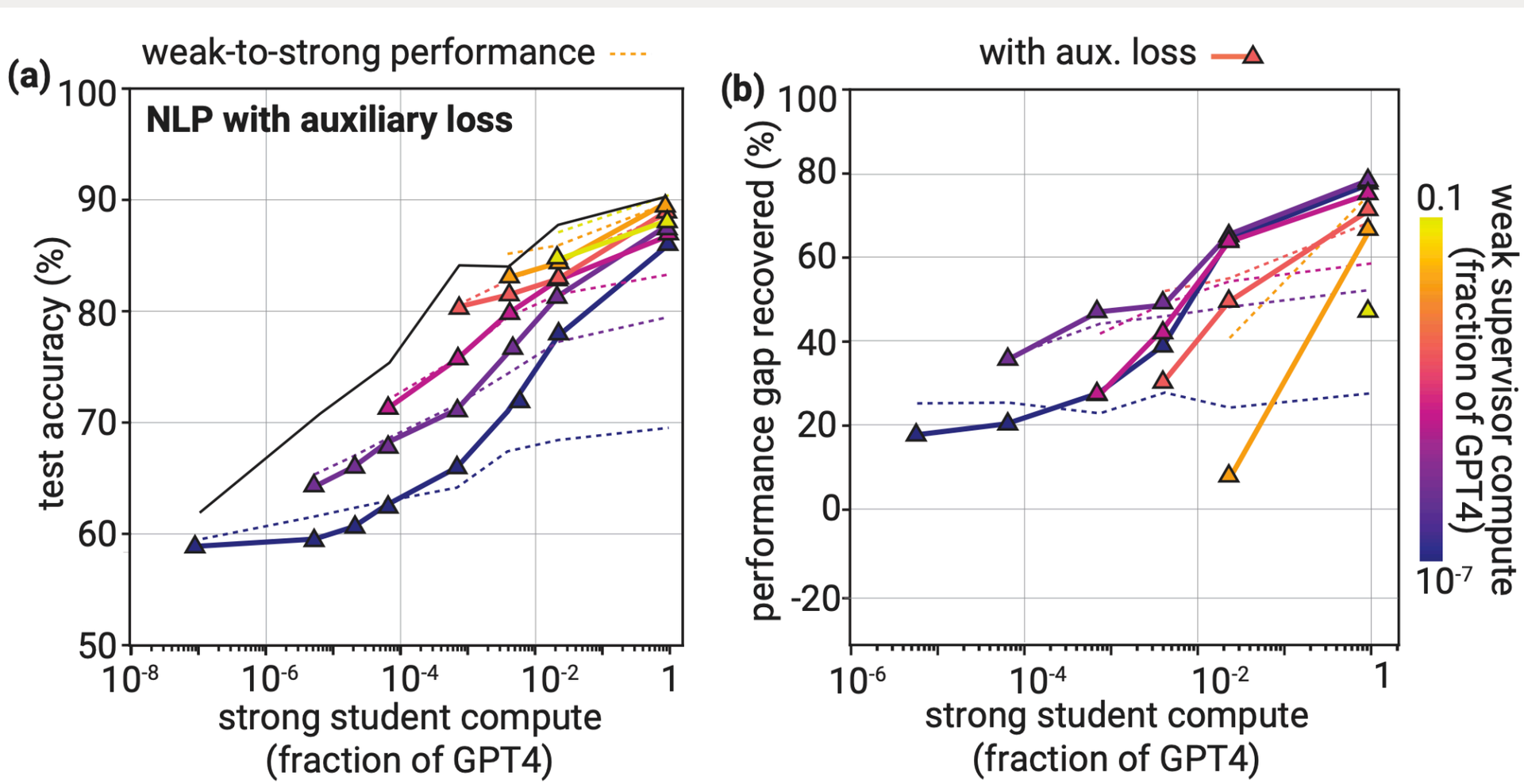


Improving Weak-to-Strong Generalization

Auxiliary Loss

- A loss term to reinforces the strong model's confidence in its own predictions.

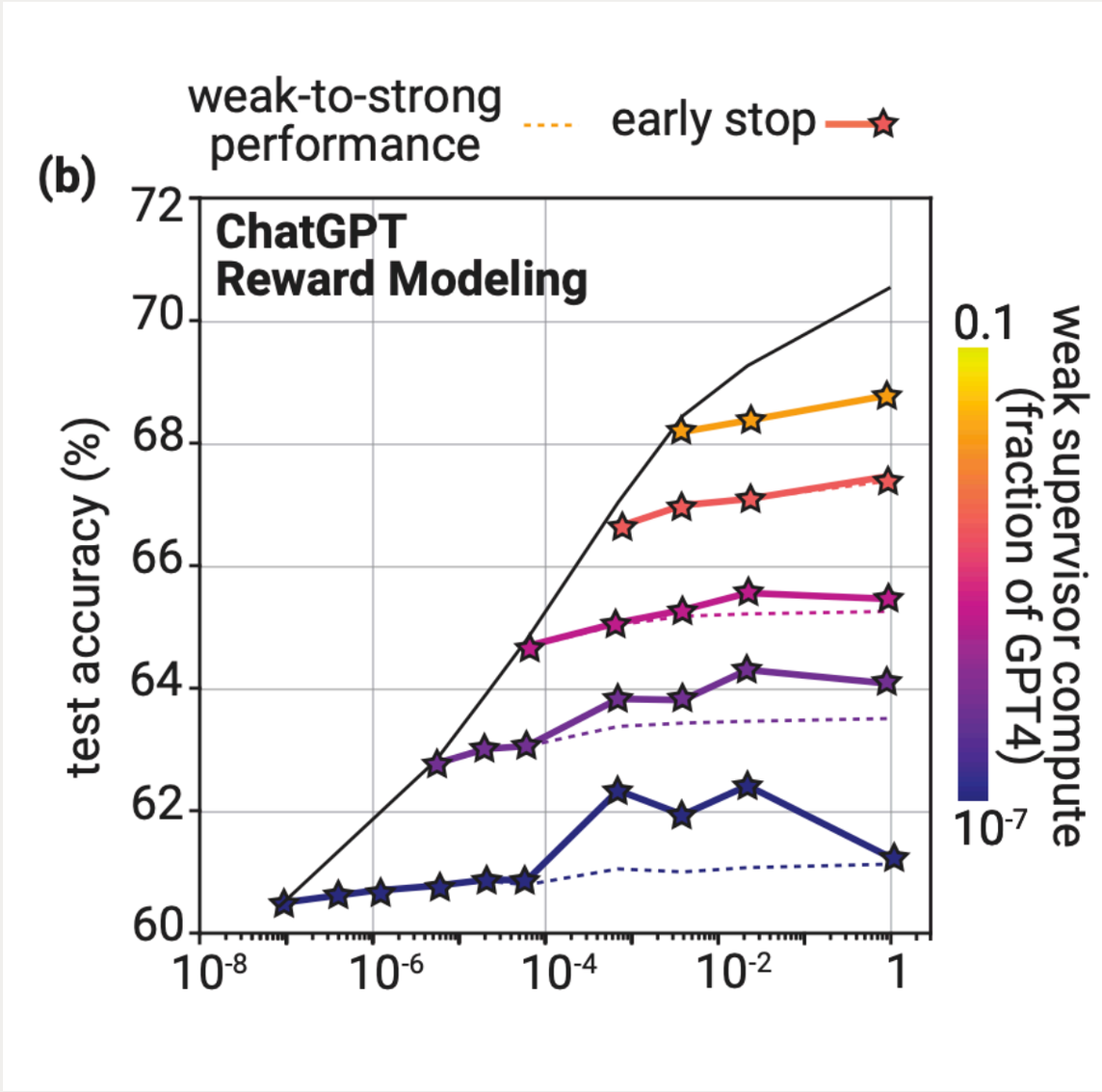
$$L_{\text{conf}}(f) = (1-\alpha)\cdot \text{CE}(f(x), f_w(x)) + \alpha \cdot \text{CE}(f(x), \hat{f}_t(x))$$



Understanding Weak-to-Strong Generalization

Weak Imitation

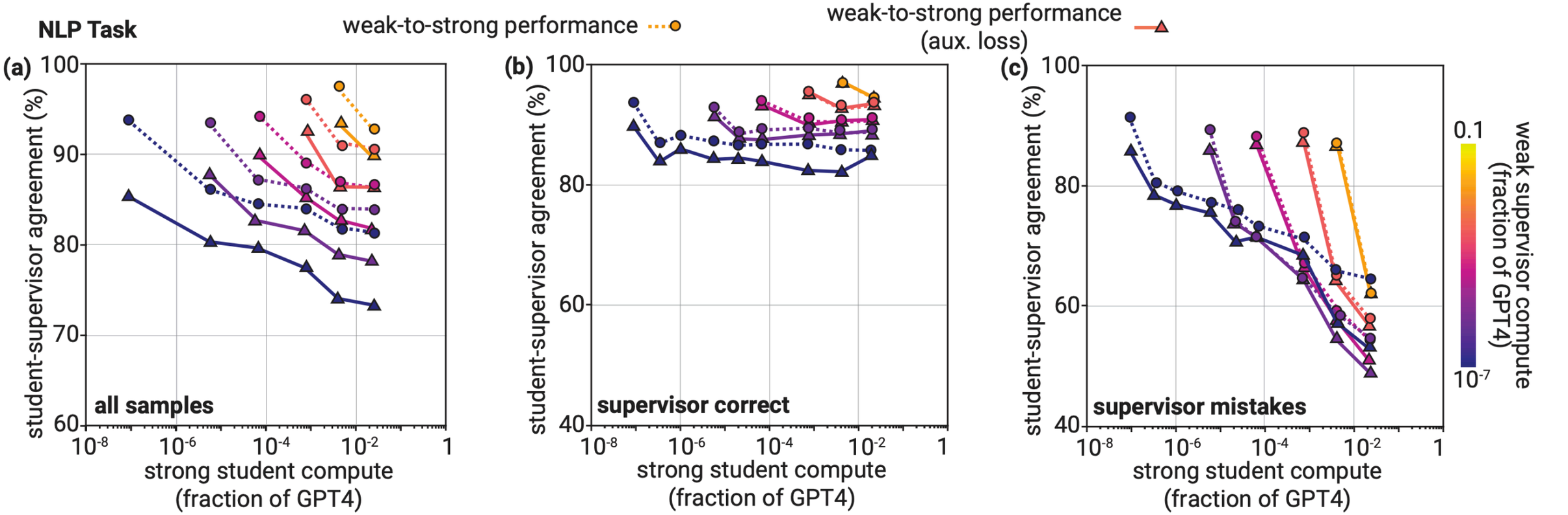
- Early Stopping



Understanding Weak-to-Strong Generalization

Weak Imitation

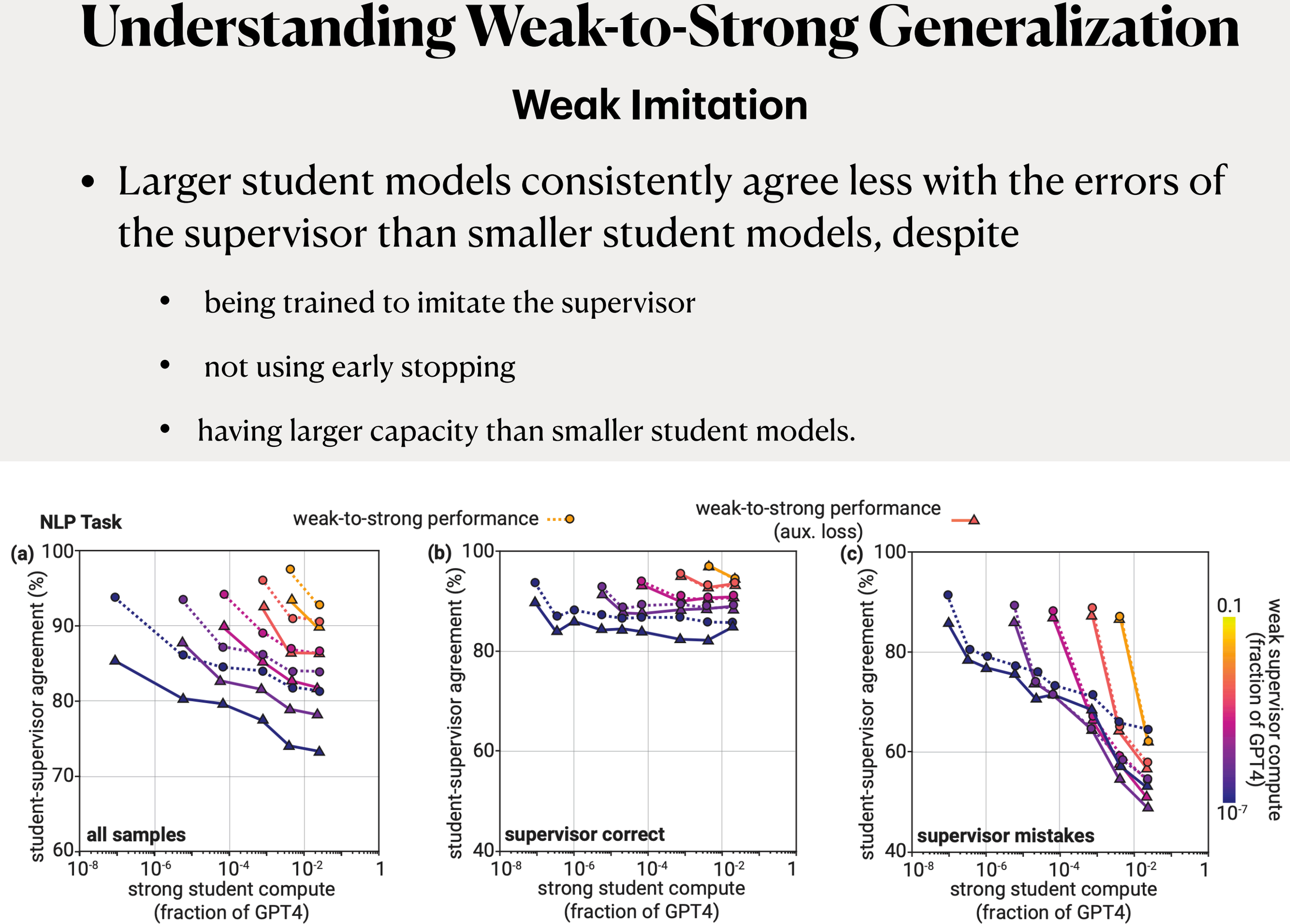
- Aux Loss prevents too much agreement



Understanding Weak-to-Strong Generalization

Weak Imitation

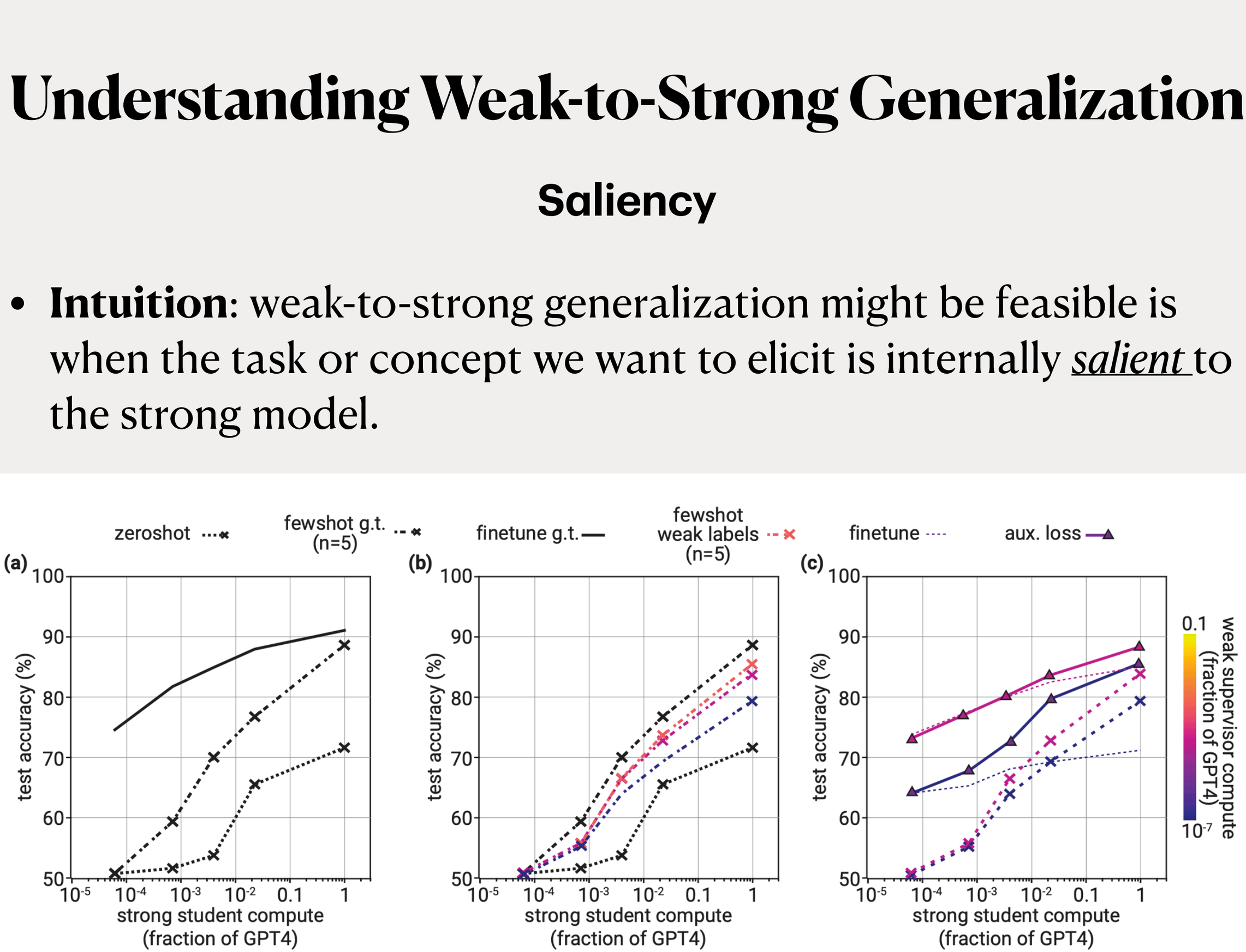
- Larger student models consistently agree less with the errors of the supervisor than smaller student models, despite
    - being trained to imitate the supervisor
    - not using early stopping
    - having larger capacity than smaller student models.



Understanding Weak-to-Strong Generalization

Saliency

- **Intuition**: weak-to-strong generalization might be feasible is when the task or concept we want to elicit is internally *salient* to the strong model.



G.T. samples for prompting

Weak samples for prompting

W2S for prompting

FT is better than prompting

Understanding Weak-to-Strong Generalization

Saliency

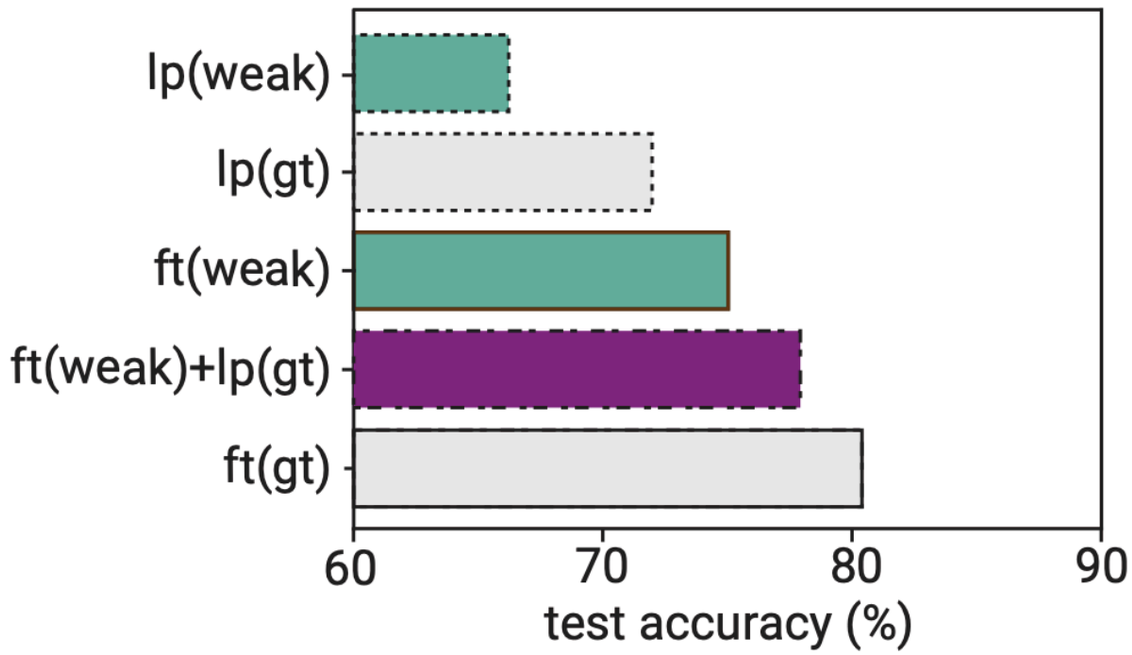


Figure 11: **Finetuning on weak supervisor labels makes the desired generalization more linearly represented.** We plot test accuracy for five different strategies, averaged across a subset of NLP tasks. **lp(weak)**: training a linear probe on the base model using weak labels, **lp(gt)**: training a linear probe on the base models using ground truth labels, **ft(weak)**: finetuning the model on weak labels, **ft(weak) + lp(gt)**: finetuning the model on weak labels *then* training a linear probe on ground truth labels, **ft(gt)**: finetuning the model on ground truth labels. Finetuning on the weak labels significantly increases the linearity of the ground truth concept.

Let’s now see some theory.

Random Features Model

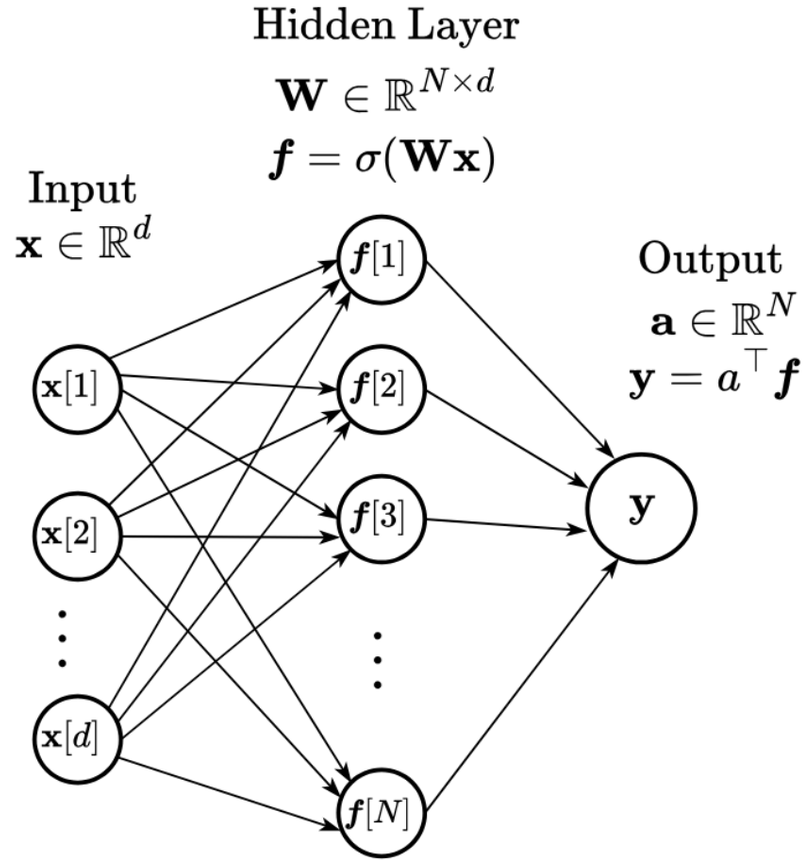
Weak-to-Strong Generalization Even in Random Feature Networks, Provably

Marko Medvedev[1,*], Kaifeng Lyu[2,*], Dingli Yu[3], Sanjeev Arora[4], Zhiyuan Li[5], Nathan Srebro[5]

[1]University of Chicago, [2]Simons Institute, University of California, [3]Microsoft Research, [4]Princeton University, [5]Toyota Technological Institute at Chicago

Random Features Model

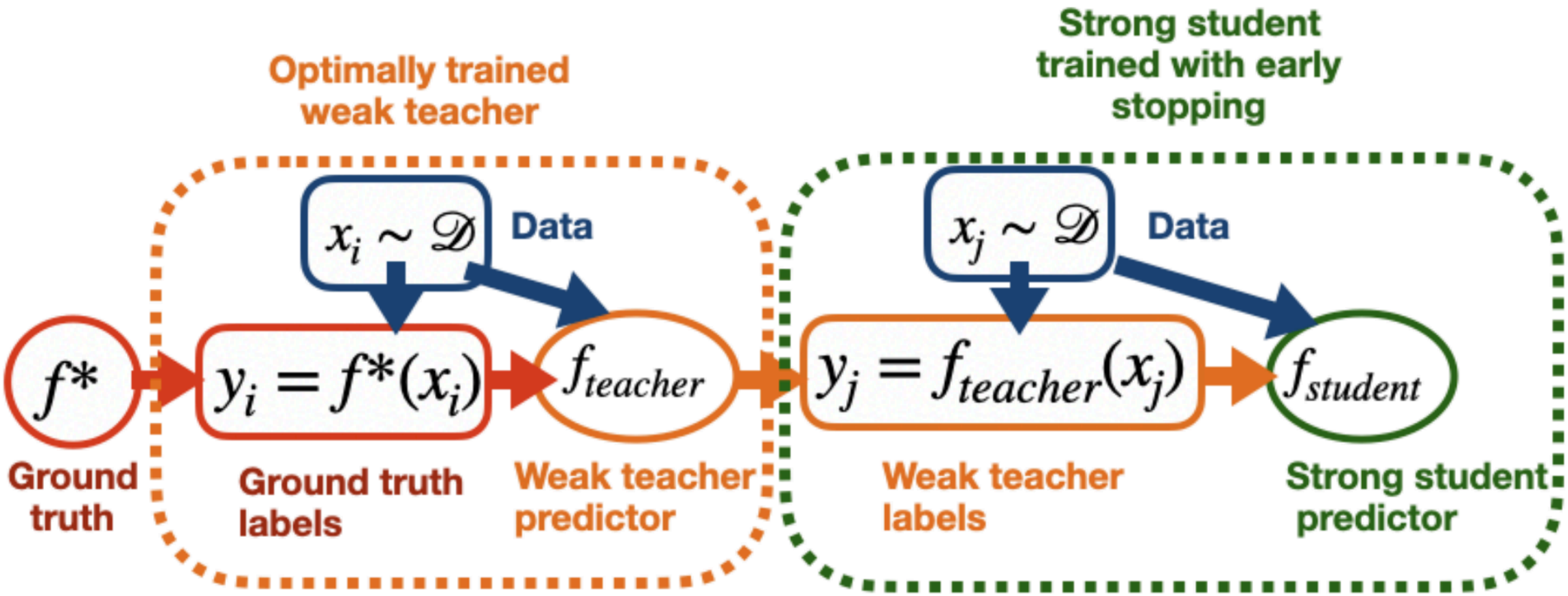
- Two-Layer fully connected neural network $f_{\mathrm{NN}}(\boldsymbol{x}) = \mathbf{a}^\top \sigma(\mathbf{W}\boldsymbol{x})$
- Random first layer weights
- Trained second layer
- Captures various deep learning phenomena.



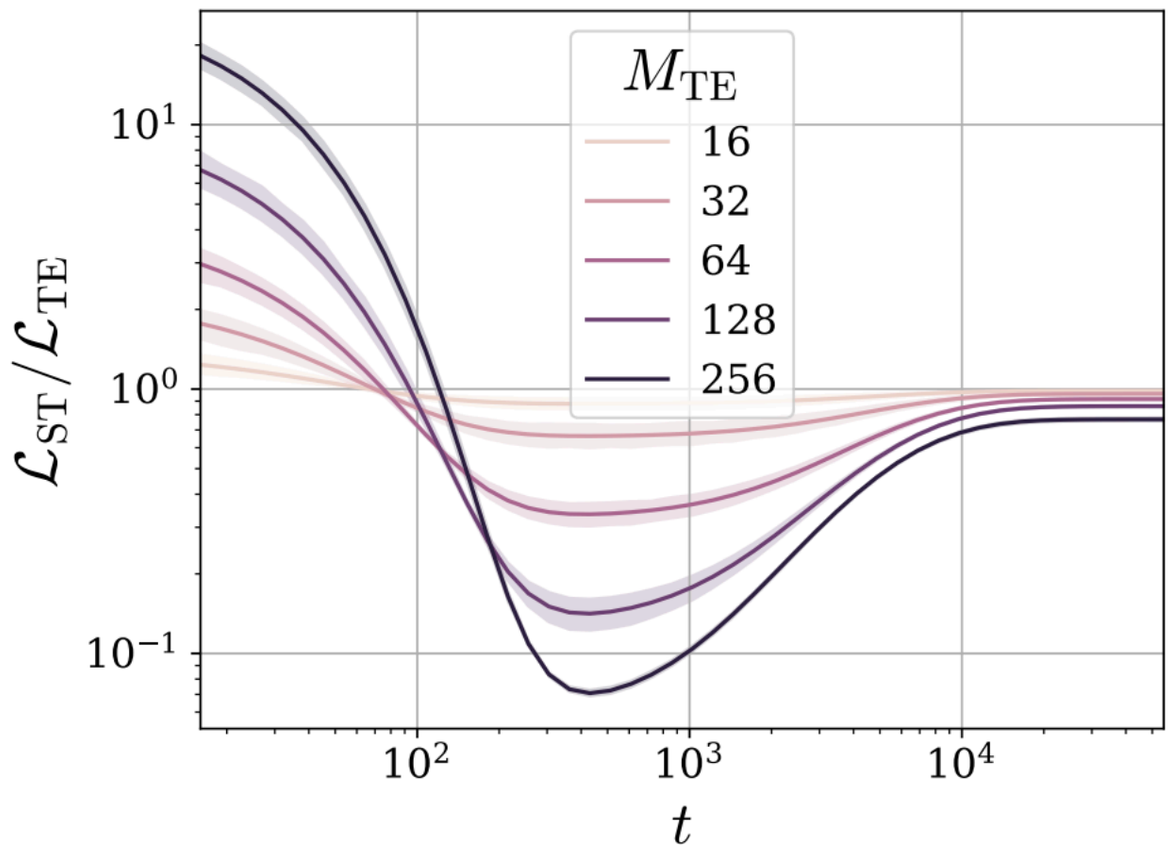
Rahimi and Recht, Random Features for Large-Scale Kernel Machines, NeurIPS 2007.

Random Features Model

- Weak model has significantly less random features compared to the strong model.
- We fit the second layer (i.e., linear probing) using gradient flow.



Random Features Model



(a) $\mathcal{L}_{\mathrm{ST}}/\mathcal{L}_{\mathrm{TE}}$ v.s. t

Early Stopping is the key here

The Role of Representations

Representations Shape Weak-to-Strong Generalization: Theoretical Insights and Empirical Predictions

Yihao Xue [1] **Jiping Li** [2] **Baharan Mirzasoleiman** [1]

The Role of Representations

- Again, assume that we have a fixed set of features and we do linear probing.

$$f_{\mathrm{w}} = \arg\min_{f \in \mathcal{F}_{\mathrm{w}}} \left(\frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} \left(f(h_{\mathrm{w}}(\tilde{\boldsymbol{x}}_i)) - \tilde{y}_i \right)^2 + \beta_{\mathrm{w}} R(f) \right).$$

$$f_{\mathrm{w2s}} = \arg\min_{f \in \mathcal{F}_{\mathrm{s}}} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \left(f(h_{\mathrm{s}}(\hat{\boldsymbol{x}}_i)) - f_{\mathrm{w}}(h_{\mathrm{w}}(\hat{\boldsymbol{x}}_i)) \right)^2 + \beta_{\mathrm{s}} R(f) \right).$$

$$f_{\mathrm{sc}} = \arg\min_{f \in \mathcal{F}_{\mathrm{s}}} \left(\frac{1}{\hat{n}} \sum_{i=1}^{\hat{n}} \left(f(h_{\mathrm{s}}(\hat{\boldsymbol{x}}_i)) - \hat{y}_i \right)^2 + R_{\mathrm{s}}(f) \right).$$

- **Prediction Gap**:

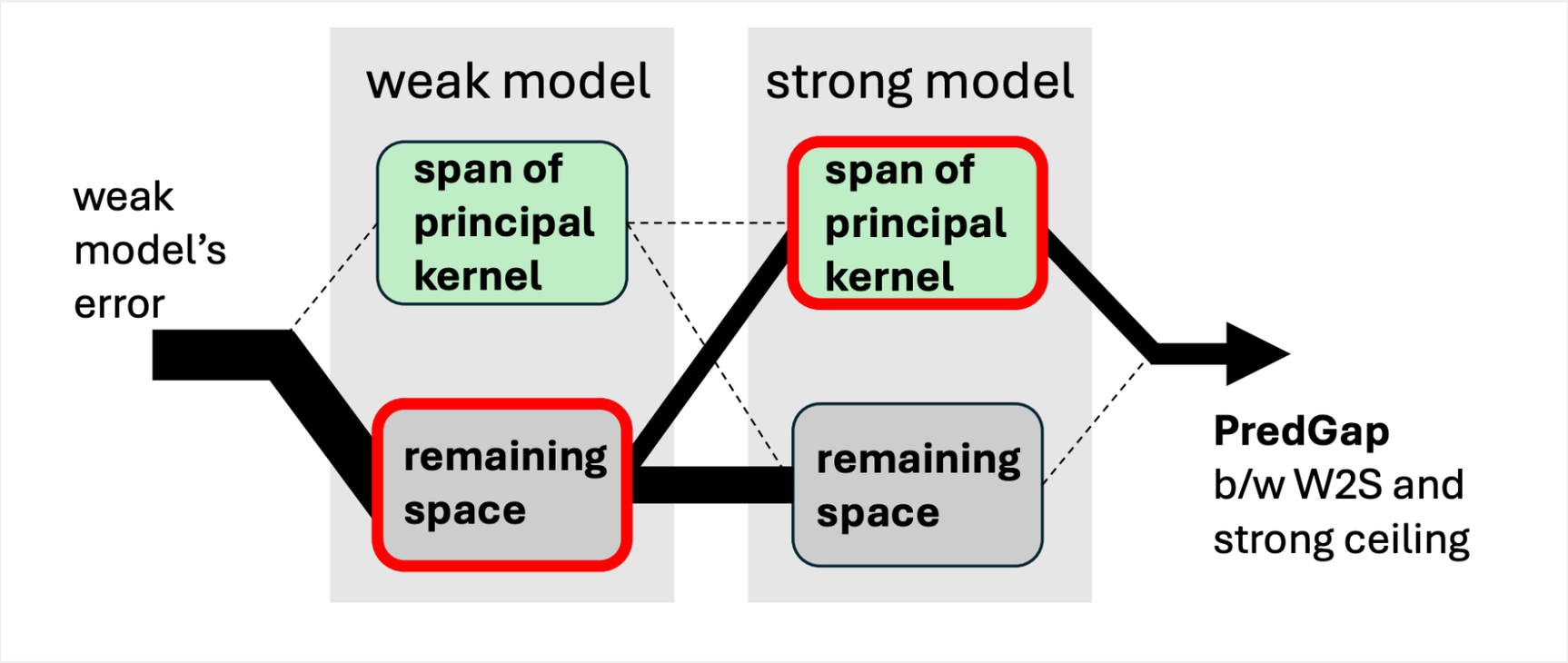
$$\textbf{PredGap} = \mathbb{E}_{(\boldsymbol{x}, y) \sim \mathcal{D}} \left[\left(f_{\mathrm{w2s}}(h_{\mathrm{s}}(\boldsymbol{x})) - f_{\mathrm{sc}}(h_{\mathrm{s}}(\boldsymbol{x})) \right)^2 \right].$$

Effect of Features

Theorem 3.8 (Main result)**.** *Under Assump. 3.7, and assuming reasonable regularization:* $\delta_w \le \beta_w = O(1)$ *and* $\delta_s \le \beta_s = O(1)$*, let* $\hat{\boldsymbol{y}} = [\hat{y}_1\ \hat{y}_2\ \dots\ \hat{y}_{\hat{n}}]^\top$*. Then, w.h.p., we have*

$$\textbf{\textit{PredGap}} = \left\| \boldsymbol{P}_s(\boldsymbol{I} - \boldsymbol{P}_w)\frac{1}{\sqrt{\hat{n}}}\hat{\boldsymbol{y}} \right\|^2 \pm o(1) \tag{1}$$

$P_s(I - P_w)$ captures "what the weak model is unable to learn but is learnable by the strong model using their respective principal representations". Therefore, it determines the mistakes that will be learned by the strong model, as discussed in the intuition. A more powerful weak model has a $\boldsymbol{P}_w$ that covers more space, shrinking $\boldsymbol{P}_s(\boldsymbol{I} - \boldsymbol{P}_w)$ and potentially leading to a smaller **PredGap**.



Discrepancies are Virtue

Discrepancies are Virtue: Weak-to-Strong Generalization through Lens of Intrinsic Dimension

Yijun Dong[1] Yicheng Li[1] Yunai Li[2] Jason D. Lee[3] Qi Lei[1]

[1]New York University [2]Shanghai Jiaotong University [3]Princeton University

{yd1319,yl9315}@nyu.edu liyunai_8528@sjtu.edu
jasonlee@princeton.edu ql518@nyu.edu

Discrepancies are Virtue

(a) **Weak teacher model** $f_w(\mathbf{x}) = \phi_w(\mathbf{x})^\top \boldsymbol{\theta}_w$ is supervised finetuned over $\widetilde{\mathcal{S}}$:

$$\boldsymbol{\theta}_w = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\| \widetilde{\boldsymbol{\Phi}}_w \boldsymbol{\theta} - \widetilde{\mathbf{y}} \right\|_2^2 + \alpha_w \left\| \boldsymbol{\theta} \right\|_2^2 . \tag{2}$$

(b) **W2S model** $f_{\mathrm{w2s}}(\mathbf{x}) = \phi_s(\mathbf{x})^\top \boldsymbol{\theta}_{\mathrm{w2s}}$ is finetuned over the strong feature ϕ_s through $\mathcal{S}_x$ and their pseudo-labels generated by the weak teacher model:

$$\boldsymbol{\theta}_{\mathrm{w2s}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\| \boldsymbol{\Phi}_s \boldsymbol{\theta} - \boldsymbol{\Phi}_w \boldsymbol{\theta}_w \right\|_2^2 + \alpha_{\mathrm{w2s}} \left\| \boldsymbol{\theta} \right\|_2^2 \tag{3}$$

(c) **Strong SFT model** $f_s(\mathbf{x}) = \phi_s(\mathbf{x})^\top \boldsymbol{\theta}_s$ is a strong baseline where the strong feature ϕ_s is supervisedly finetuned over the small labeled set $\widetilde{\mathcal{S}}$ directly:

$$\boldsymbol{\theta}_s = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\| \widetilde{\boldsymbol{\Phi}}_s \boldsymbol{\theta} - \widetilde{\mathbf{y}} \right\|_2^2 + \alpha_s \left\| \boldsymbol{\theta} \right\|_2^2 . \tag{4}$$

(d) **Strong ceiling model** $f_c(\mathbf{x}) = \phi_s(\mathbf{x})^\top \boldsymbol{\theta}_c$ is a reference for the ceiling performance where ϕ_s is supervisedly finetuned over $\mathcal{S} \cup \widetilde{\mathcal{S}}$, assuming access to the unknown labels $\mathbf{y} = [y_1, \ldots, y_N]^\top$:

$\boxed{\alpha_w, \alpha_{\mathrm{w2s}}, \alpha_s, \alpha_c \to 0}$

$$\boldsymbol{\theta}_c = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \left\| \begin{bmatrix} \widetilde{\boldsymbol{\Phi}}_s \\ \boldsymbol{\Phi}_s \end{bmatrix} \boldsymbol{\theta} - \begin{bmatrix} \widetilde{\mathbf{y}} \\ \mathbf{y} \end{bmatrix} \right\|_2^2 + \alpha_c \left\| \boldsymbol{\theta} \right\|_2^2 . \tag{5}$$

Discrepancies are Virtue

Theorem 1 (W2S model (formally in B.1)). *Assuming Assumptions 1 and 2 and* $\phi_w(\mathbf{x}) \sim \mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma}_w)$, *for* $n > d_w + 1$, $\mathbf{ER}(f_{\mathrm{w2s}}) = \mathbf{Var}(f_{\mathrm{w2s}}) + \mathbf{Bias}(f_{\mathrm{w2s}})$ *satisfies*

$$\mathbf{Var}(f_{\mathrm{w2s}}) = \frac{\sigma^2}{n - d_w - 1}\left(d_{s\wedge w} + \frac{d_s}{N}(d_w - d_{s\wedge w})\right),$$

$$\mathbf{Bias}(f_{\mathrm{w2s}}) \leqslant \frac{\rho_w(n)}{n} + \frac{\rho_s(N)}{N} \leqslant \rho_w + \rho_s,$$ = zero; variance dominated regime

where the inequality for $\mathbf{Bias}(f_{\mathrm{w2s}})$ *is strict if* $\rho_w(n)/n > 0$ *and* $d_s < d_w$.

$$\mathbf{Var}(f_w) = \sigma^2 \frac{d_w}{n}, \quad \mathbf{Bias}(f_w) = \frac{\rho_w(n)}{n} \leqslant \rho_w,$$

$$\mathbf{Var}(f_s) = \sigma^2 \frac{d_s}{n}, \quad \mathbf{Bias}(f_s) = \frac{\rho_s(n)}{n} \leqslant \rho_s,$$

$$\mathbf{Var}(f_c) = \sigma^2 \frac{d_s}{N+n}, \quad \mathbf{Bias}(f_c) = \frac{\rho_s(N+n)}{N+n} \leqslant \rho_s,$$

$$\rho_s = \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[(\phi_s(\mathbf{x})^\top\boldsymbol{\theta} - f_*(\mathbf{x}))^2\right],$$

$$\rho_w = \min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[(\phi_w(\mathbf{x})^\top\boldsymbol{\theta} - f_*(\mathbf{x}))^2\right],$$

Discrepancies are Virtue

- **Intuition.**
- **Task:** classify the brand of a car in an image
- **Strong Features**: information in the logo
  (low intrinsic dimension ds)
- **Weal Features**: complex information in the design
  (large intrinsic dimension dw).

Quantifying the Gains

Quantifying the Gain in Weak-to-Strong Generalization

Moses Charikar
Stanford University
moses@cs.stanford.edu

Chirag Pabbaraju
Stanford University
cpabbara@cs.stanford.edu

Kirankumar Shiragur
Microsoft Research
kshiragur@microsoft.com

Quantifying the Gains

Main Result

Let $\text{MSE}(p, q) = \mathbb{E}_x[p(x) - q(x)]^2$ and suppose the strong model obtains f_{sw} from a convex space $\boldsymbol{\mathcal{F}}_s$

$$f_{sw} = \text{argmin}_{f \in \boldsymbol{\mathcal{F}}_s} \text{MSE}(f \circ h_s, f^* \circ h^*)$$

Then, assuming realizability (i.e., $f^* \circ h^* = f_s \circ h_s$ for some $f_s \in \boldsymbol{\mathcal{F}}_s$),

$$\underbrace{\text{MSE}(f_w \circ h_w, f^* \circ h^*) - \text{MSE}(f_{sw} \circ h_s, f^* \circ h^*)}_{\text{GAIN IN PERFORMANCE}} \geq \underbrace{\text{MSE}(f_{sw} \circ h_s, f_w \circ h_w)}_{\text{MISFIT BETWEEN WEAK AND STRONG MODEL}}$$

Future Work

- **Updating the representation (beyond linear probing)**:
  The strong model can learn the features of the weak model.
  - Language learning example.
- **The role of implicit regularization:**
  Even in very simple linear models.
  - Linear regression example
- Stay tuned :)

Thanks!