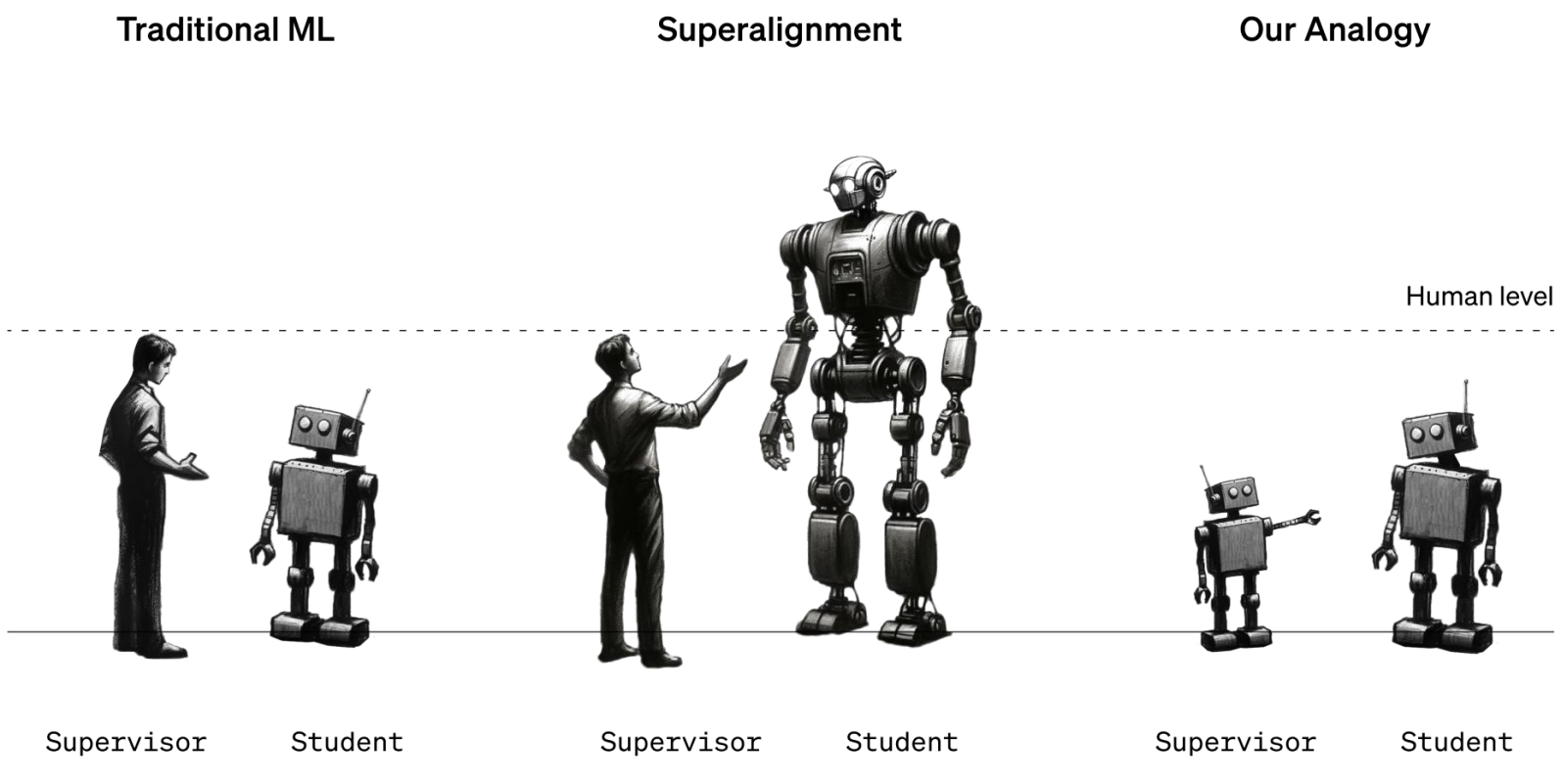
Mian Zhong & Milton Lin

How can weak supervisors control strong models?

Motivation from authors

- **Superalignment**
  The challenge of weak human supervision for complex AI behavior.
- **Can we use weak models to elicit strong model abilities?**



Methodology

1. **Create a weak supervisor**
   Small pretrained models (e.g., GPT-2) are fine-tuned on ground truth labels to generate weak supervision for the task

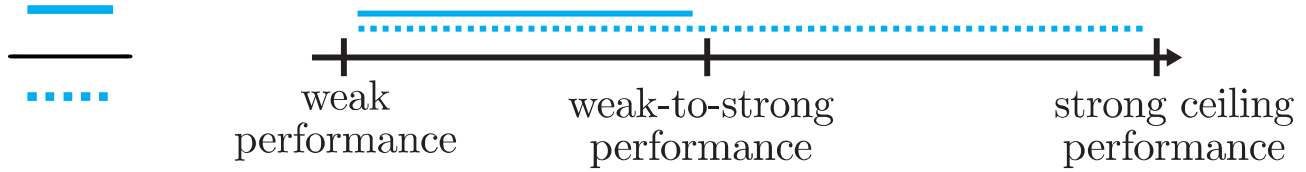
2. **Train strong models with weak supervision**
   Strong models (e.g., GPT-4) are fine-tuned using the weak labels from the supervisor in (1). In (3) we evaluate the generalization ability of the strong model, defined as *weak-to-strong performance*.
3. **Compare with Strong Ceiling Performance**
   As a baseline, the strong models are also trained using ground truth labels to establish the *strong ceiling performance*. The difference in performance between weak and strong supervision is measured using *Performance Gap Recovered (PGR)*.

Performance Gap Recovered (PGR)

PGR quantifies **how much of the gap between weak performance and strong performance is bridged by training on weak labels.**

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

Understanding PGR



- If PGR = 1:
  The strong model performs as well as it would with ground truth supervision, achieving the *strong ceiling performance*.

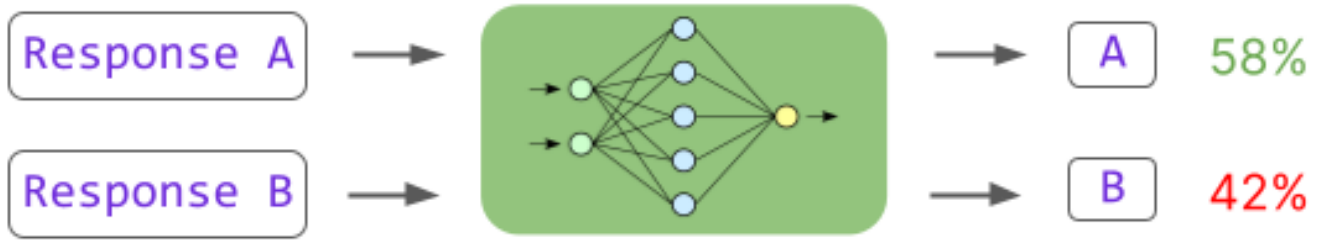
- If PGR = 0:
  The strong model does no better than the weak supervisor

- Intermediate values (0 < PGR < 1):
  Represent partial recovery, where the strong model generalizes beyond the weak labels but does not reach its full potential. A higher PGR means that weak-to-strong generalization is more successful.

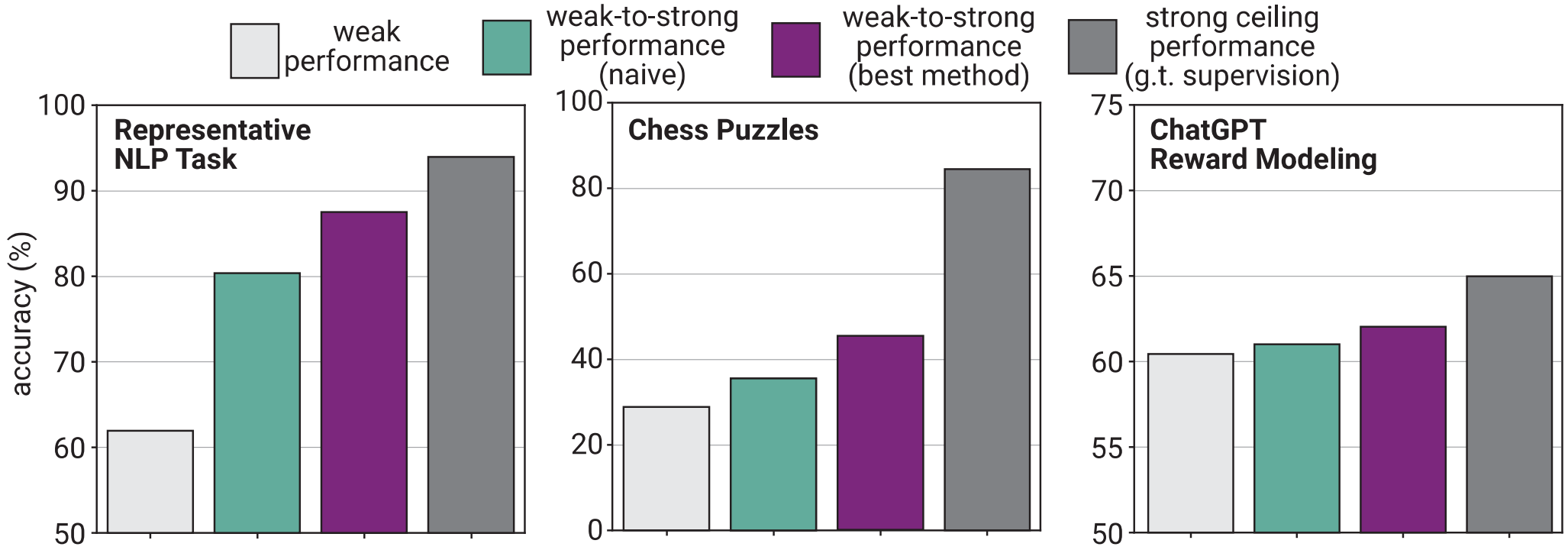
Evaluation tasks

- NLP Binary Classification
- Reward Modeling
- Chess

Dataset	Original Source
BoolQ	Clark et al. (2019)
CosmosQA	Huang et al. (2019)
DREAM	Sun et al. (2019)
ETHICS [Justice]	Hendrycks et al. (2020a)
ETHICS [Deontology]	Hendrycks et al. (2020a)
ETHICS [Virtue]	Hendrycks et al. (2020a)
ETHICS [Utilitarianism]	Hendrycks et al. (2020a)
FLAN ANLI R2	Nie et al. (2019); Wei et al. (2021)
GLUE CoLA	Warstadt et al. (2019); Wang et al. (2018)
GLUE SST-2	Socher et al. (2013); Wang et al. (2018)
HellaSwag	Zellers et al. (2019)
MCTACO	Zhou et al. (2019)
OpenBookQA	Mihaylov et al. (2018)
PAWS	Zhang et al. (2019)
QuAIL	Rogers et al. (2020)
PIQA	Bisk et al. (2020)
QuaRTz	Tafjord et al. (2019)
SciQ	Welbl et al. (2017)
Social IQa	Sap et al. (2019)
SuperGLUE MultiRC	Khashabi et al. (2018); Wang et al. (2019)
SuperGLUE WIC	Pilehvar & Camacho-Collados (2018); Wang et al. (2019)
Twitter Sentiment	Zhang et al. (2019)



Finding I: strong models consistently outperform their weak supervisors across tasks.

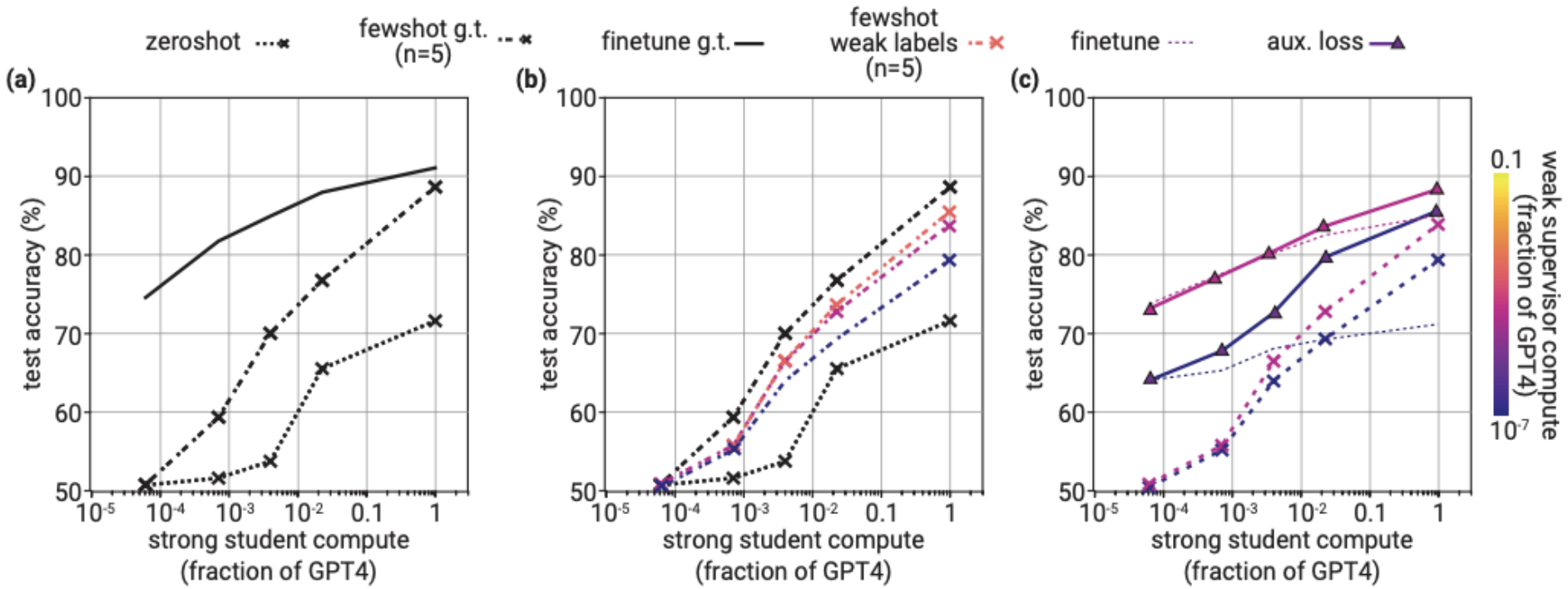


[Figure 2] In NLP tasks, fine-tuning GPT-4 on GPT-2-level labels recovers about 50% of the performance gap between the two models.

In contrast, reward modeling tasks exhibit poor generalization with a much lower PGR, even with increasing compute.

Why is weak-to-strong possible? Conjecture:

- **Latent Knowledge and Pretraining:**
  strong models leverage their pretraining knowledge to perform tasks. Weak labels serve as a signal to elicit this latent knowledge.
- **Same phenomena can be seen in prompting:** **[5.2.1, figure 7]**



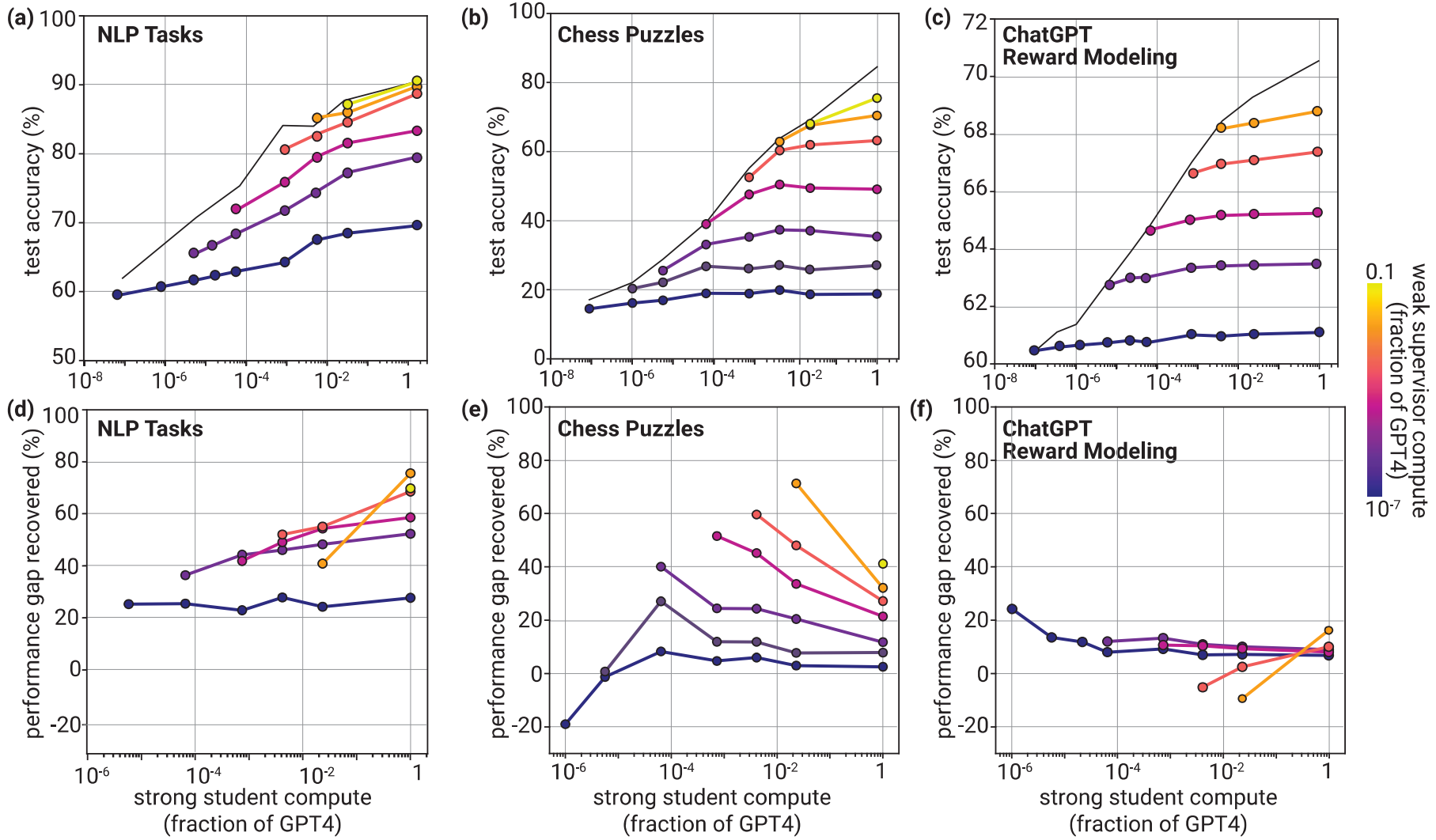
Finding II: level of generalization depends on the task.

- Weak-to-strong generalization for reward modeling is **poor** compared to NLP.
  It typically recovers only about 10% of the performance difference between the weak supervisor and the ground truth performance.

- **Saliency [5.2]** is how well a task's relevant knowledge can be elicited.
  But how to measure it? E.g. Linear representation.

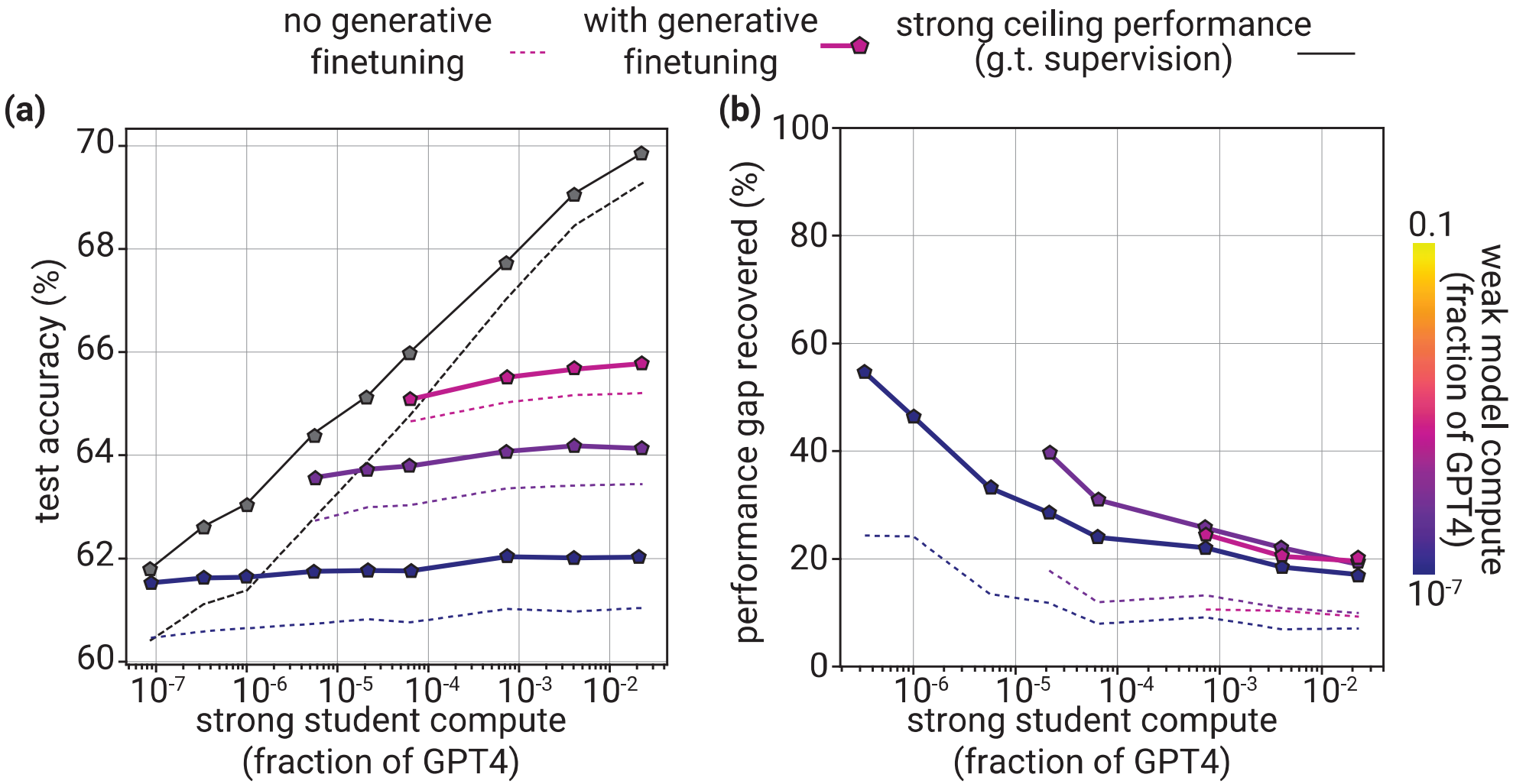
Good on NLP/♟, Poor on

[Figure 3]



Does this explain the case of reward model? Applying

generative fine-tuning to reward modeling tasks improves weak-to-strong generalization, particularly when combined with early stopping.[5.2.2]



Finding III: various methods can significantly improve the results.

- Auxiliary confidence loss : Strong models trust its own when disagreement occurs

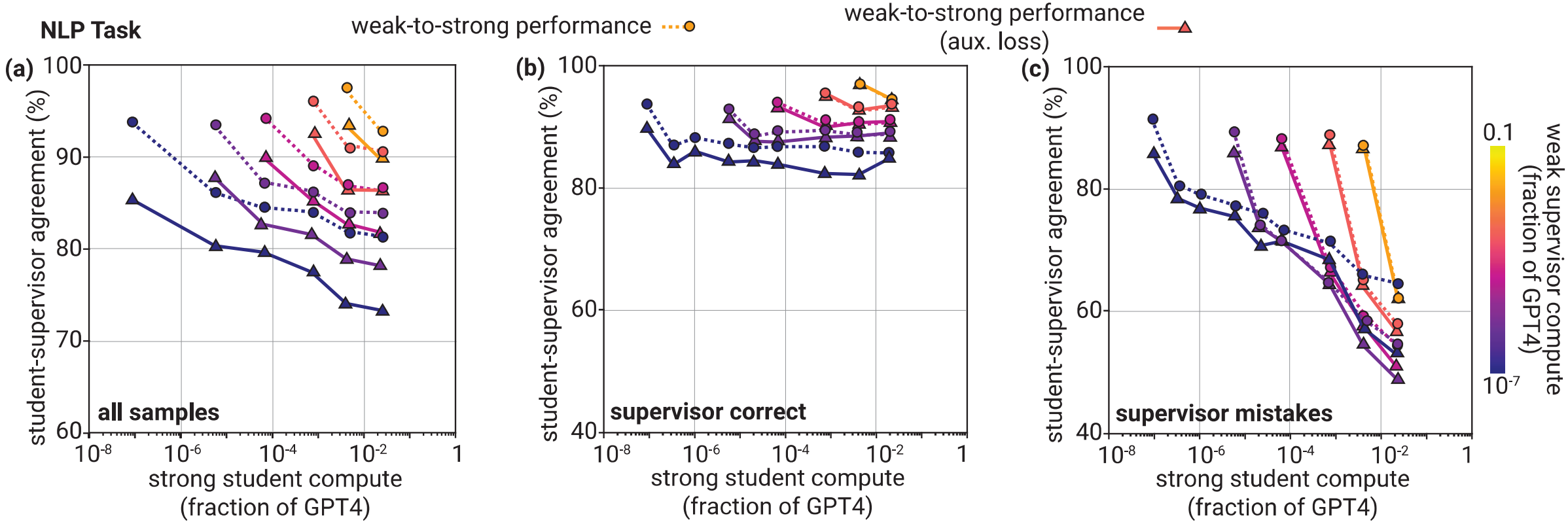
- Bootstrapping: Not helpful for NLP/RM

- Early stopping and size (next two slides)

Student-supervisor Agreement

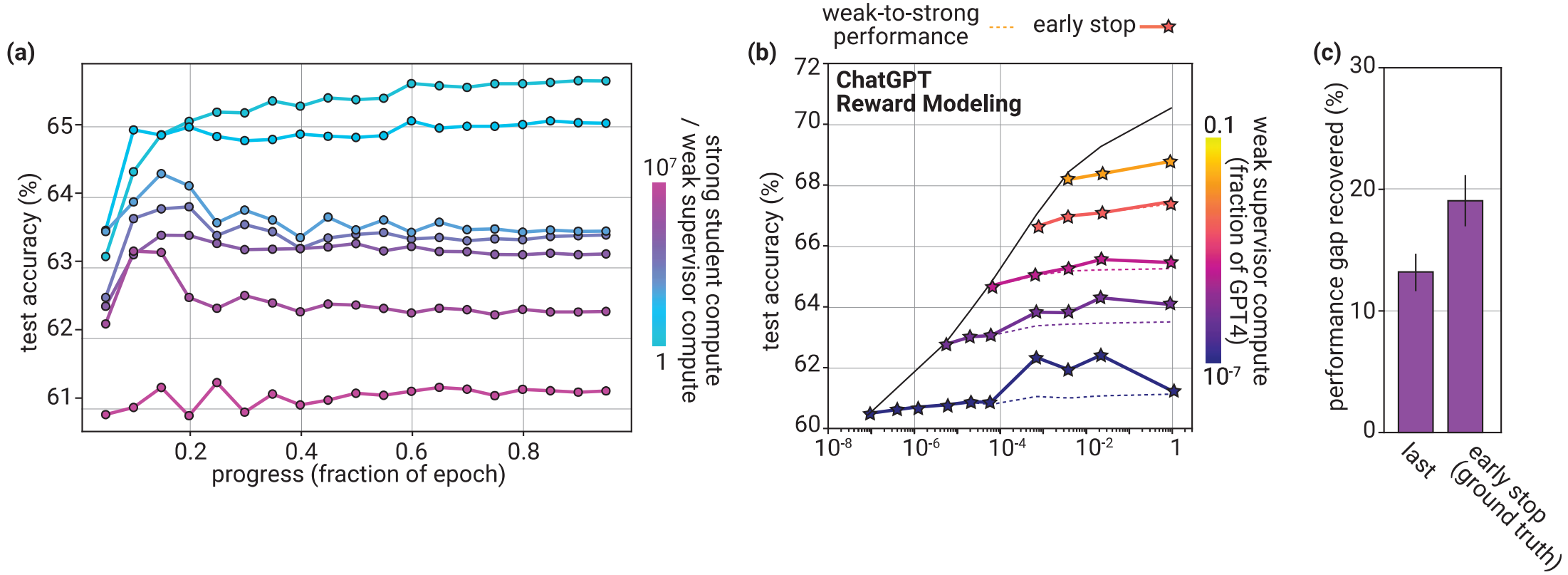
High agreement between the strong and weak models "means" the strong model is imitating the weak supervisor's predictions. [5.1.2 Figure 8]

- As strong models grow larger, they agree *less* with the weak supervisor, especially when the supervisor is wrong.



What could be a failure mode? Imitation

strong model may "overfit" to the weak supervisor's labels, meaning the strong model might imitate the weak supervisor's mistakes instead of improving upon them. [5.1]



[Figure 7] shows how performance (test accuracy) changes over the course of training (measured in fractions of an epoch). Stopping training before overfitting occurs can improve performance.

Relating back to the Alignment problem – limitations of the approach

- The setup cannot fully capture the same type of errors in superalignment
- Latent knowledge in superalignment
  - Our pretraining leakage → overly optimistic performance metrics

Other works

- Weak judge v.s. 2 strong debaters
- Improve the weak supervisor
  - Can a non-clinician elicit only appropriate medical advice with LLM?