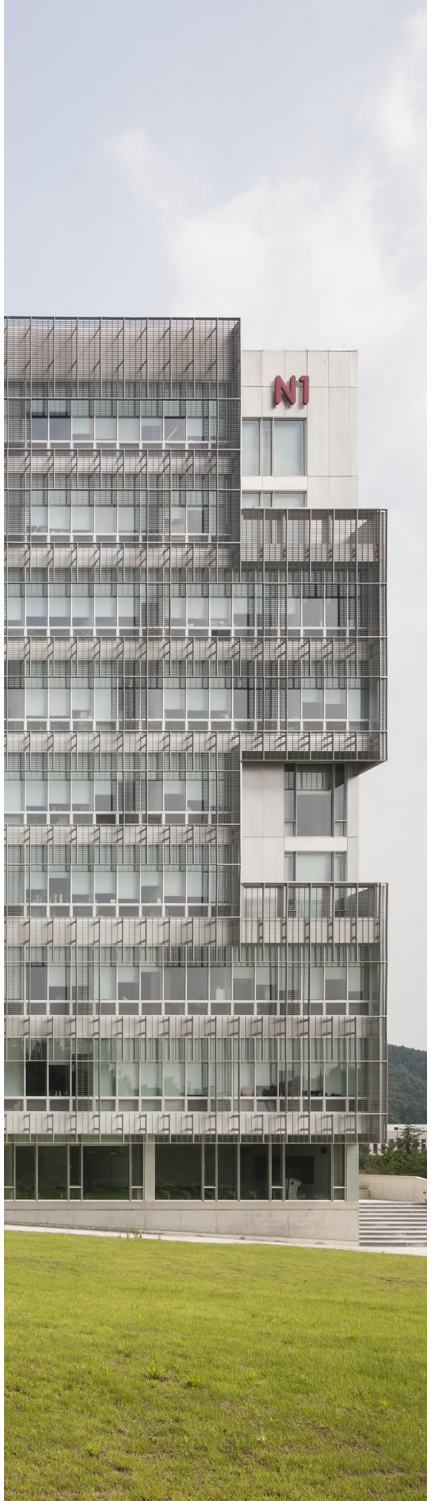


Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

ICLR 2024 (oral)

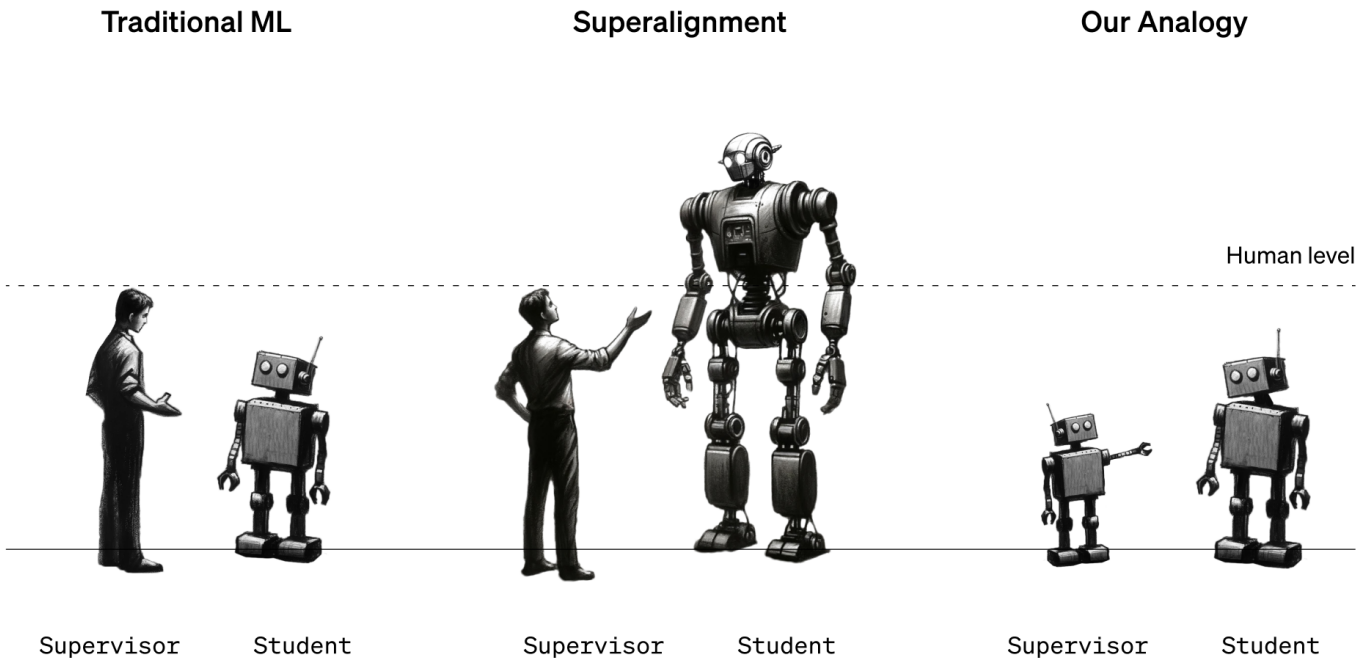
OpenAI Superalignment Generalization Team

KAIST

Overview

Superhuman model (e.g. GPT-4, Gemini, Claude)

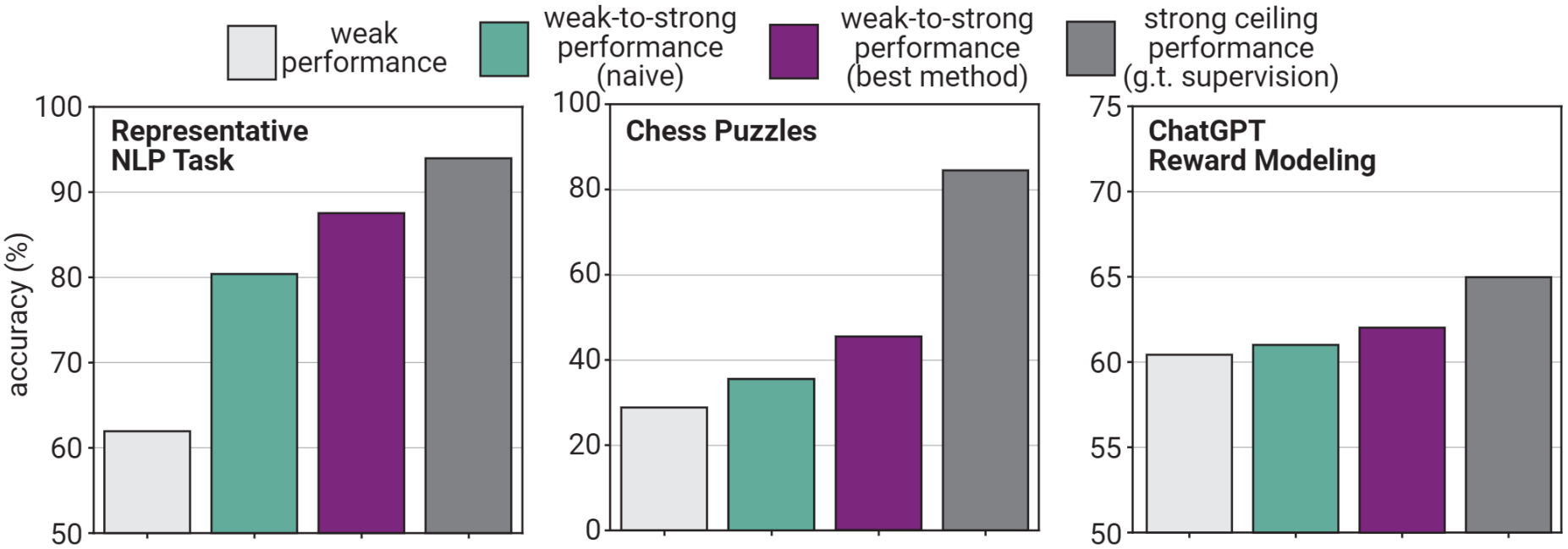
- Recent aligning method: Reinforcement Learning from Human Feedback (RLHF)
- Can humans evaluate responses from superhuman models? (e.g. millions of code)
- **GOAL**: Human supervision $\rightarrow$ Weaker model supervision



Can we use weak models to supervise strong models?

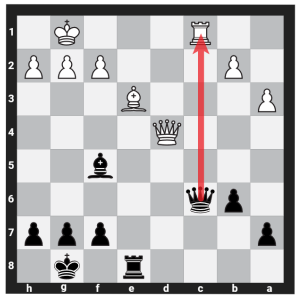
Main Contributions

1. Strong pretrained models outperform their weak supervisors. (light-gray vs. green)
2. Naively finetuning by weak supervisors is not enough. (green vs. dark-gray)
3. Improving weak-to-strong generalization is tractable. (green vs. purple)



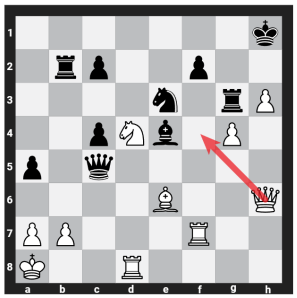
Three Types of Tasks

1. 22 Natural Language Processing tasks
2. Chess puzzles: Find the next best move



Prompt: "1. d4 1... Nf6 2. Nf3 2... d5 3. e3 3... e6 4. Bd3 4... c5 5. c3 5... Be7 6. Nbd2 6... O-O 7. O-O 7... Nc6 8. Re1 8... Bd7 9. e4 9... dxe4 10. Nxe4 10... cxd4 11. Nxf6+ 11... Bxf6 12. cxd4 12... Nb4 13. Be4 13... Qb6 14. a3 14... Nc6 15. d5 15... exd5 16. Bxd5 16... Bf5 17. Bxc6 17... Qxc6 18. Nd4 18... Bxd4 19. Qxd4 19... Rfe8 20. Rxe8+ 20... Rxe8 21. Be3 21... b6 22. Rc1 22..."

Label: " Qxc1+"



Prompt: "1. e4 1... e5 2. Nc3 2... Nf6 3. Nf3 3... Nc6 4. Bb5 4... Bc5 5. Bxc6 5... dxc6 6. d3 6... Bg4 7. h3 7... Bxf3 8. Qxf3 8... O-O 9. g4 9... Bb4 10. Bd2 10... Nd7 11. h4 11... Be7 12. g5 12... Nc5 13. O-O-O 13... Qd7 14. h5 14... Qd8 15. Qg3 15... Ne6 16. Rdg1 16... b5 17. Qxe5 17... a5 18. f4 18... Re8 19. Qf5 19... b4 20. Na4 20... Nd4 21. Qg4 21... c5 22. f5 22... Ra6 23. f6 23... Bd6 24. fxg7 24... Kxg7 25. Rg2 25... Qc8 26. h6+ 26... Kg8 27. Qh5 27... Qd7 28. Rf1 28... Re6 29. Rgf2 29... Rg6 30. c3 30... bxc3 31. Nxc3 31... a4 32. Nd5 32... Qb5 33. Nf6+ 33... Kh8 34. Qh3 34... Rb6 35. Be3 35... Ne6 36. Nxh7 36... Qxd3 37. Rd1 37... Qc4+ 38. Kb1 38... Qxe4+ 39. Ka1 39... Be5 40. Nf6 40... Qc4 41. Nd5 41... Rb7 42."

Label: " Qf5"

3. ChatGPT Reward Modeling (d, c_1, c_2, y)
   - d: dialog between a user and an assistant
   - c_1 and c_2: pair of completions for d
   - y: label, 1 if labeler preferred c_2, 0 otherwise

Methodology and Measures

- Strong student model: A family of GPT-4
- Weak teacher model: GPT-2-level model for NLP and chess, GPT-3.5-level model for RM

Weak-to-Strong model

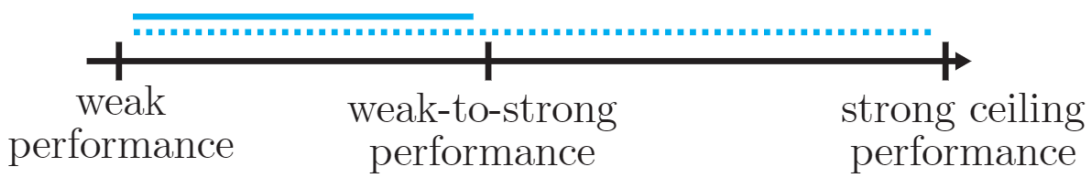
1. **Split the dataset**: one for training a weak model and the other for generating *weak labels*.
2. **Train the weak model** to generate *weak labels*.
3. **Train a strong model** using *weak labels*.

Strong-ceiling model

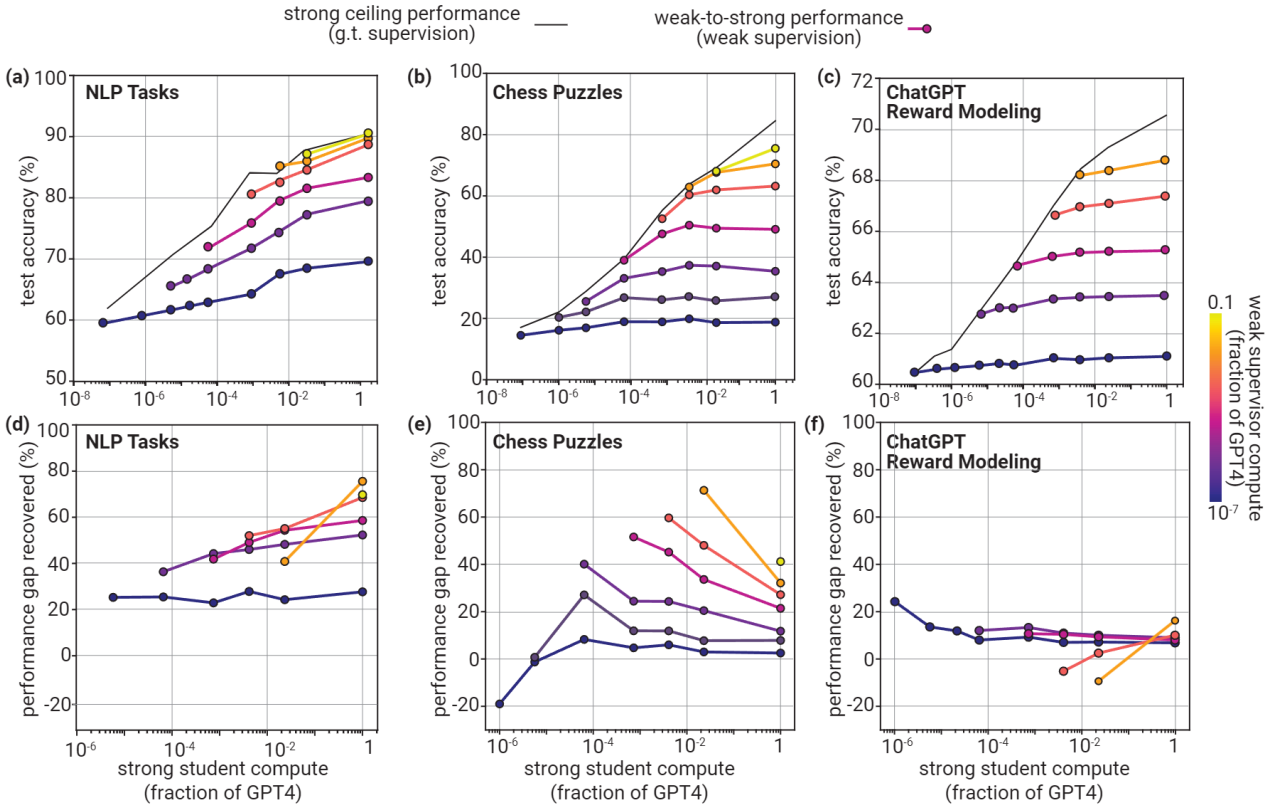
1. Train strong model using *ground truth labels*.

Performance gap recovered (PGR)

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

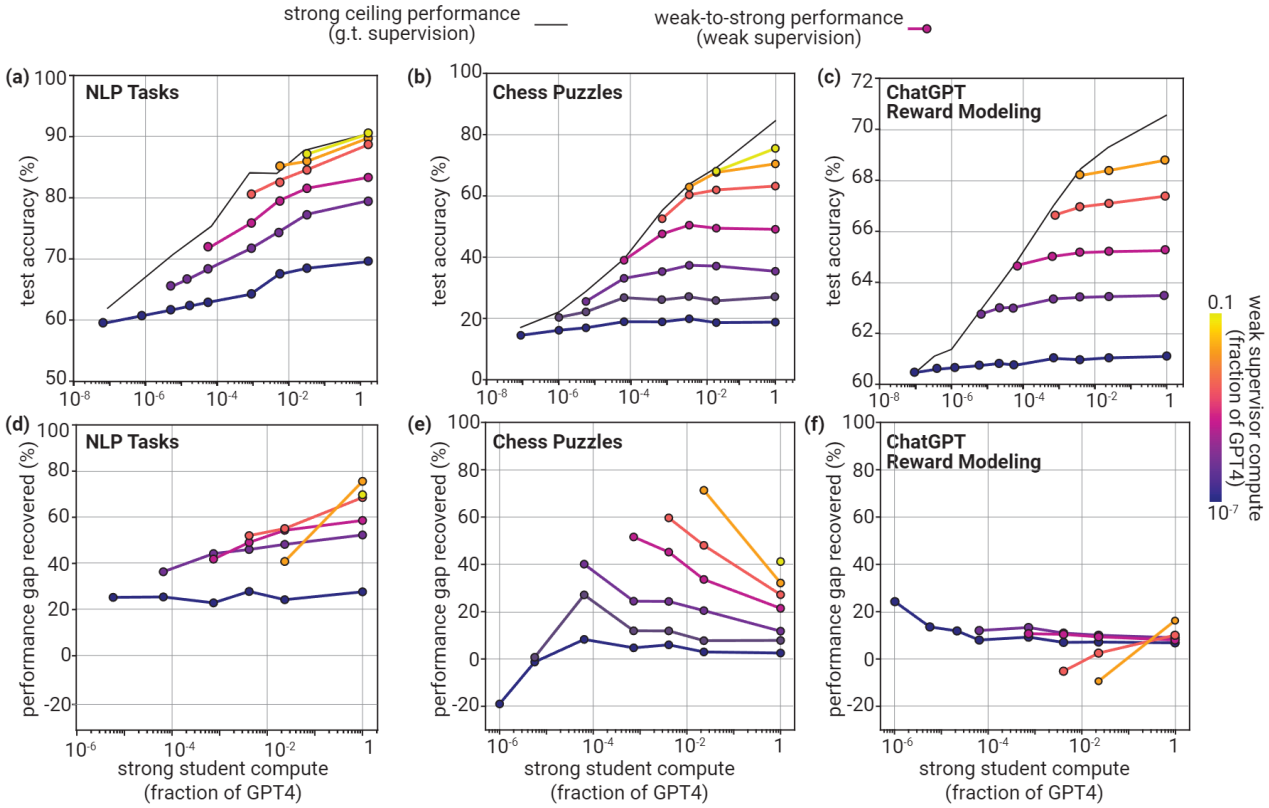


Naive Finetuning Results



- x-axis: strong student model size, colors: weak teacher model size
- black line: strong-ceiling performance, color line: weak-to-strong performance
- (a), (b), (c): test accuracy, (d), (e), (f): PGR

Naive Finetuning Results

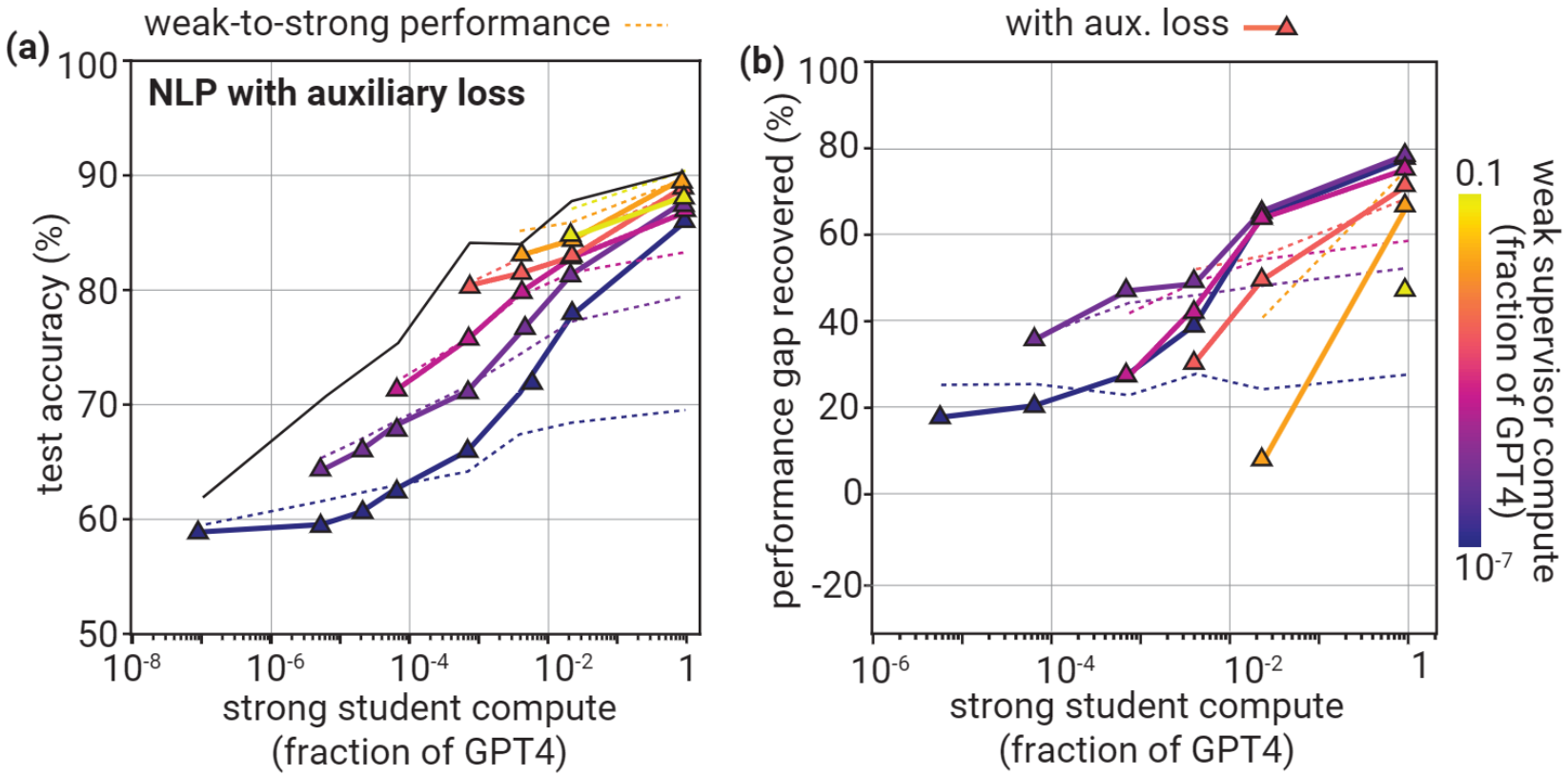


1. **PGRs are universally positive**: strong student model > weak teacher model
2. **NLP Tasks**: PGR > 20%, often above 50%.
3. **Chess Puzzles**: PGR decreases with the student size.
4. **GhatGPT Reward Modeling**: PGR never exceeds 20%.

Improving Weak-to-Strong Generalization

Confidence Loss

- **Idea**: reinforces the strong model's confidence in its own predictions, even they disagree with weak labels. cf.) conditional entropy minimization
- Dramatically improves generalization gaps in **NLP tasks**.

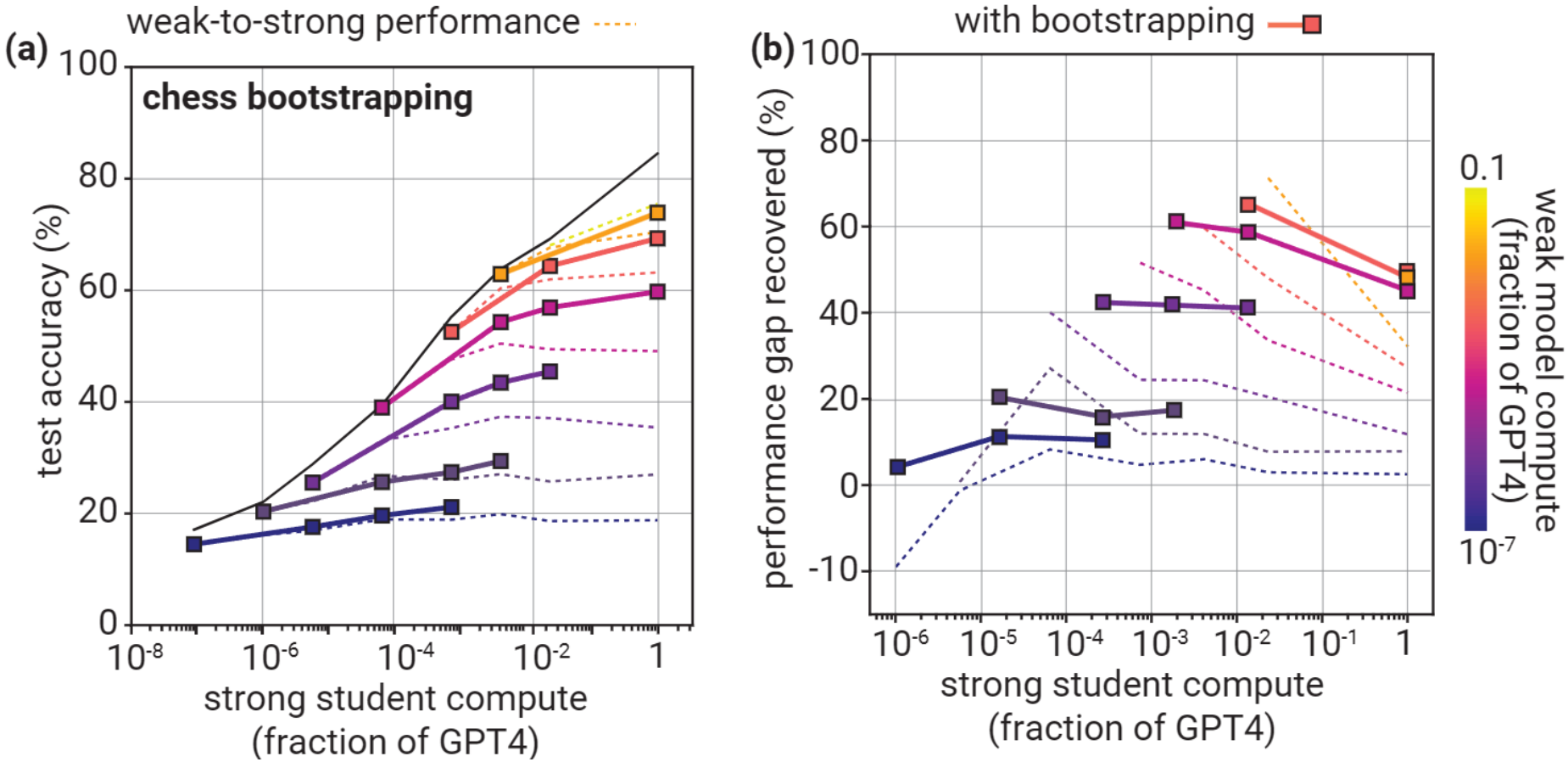


KAIST

Improving Weak-to-Strong Generalization

Bootstrapping

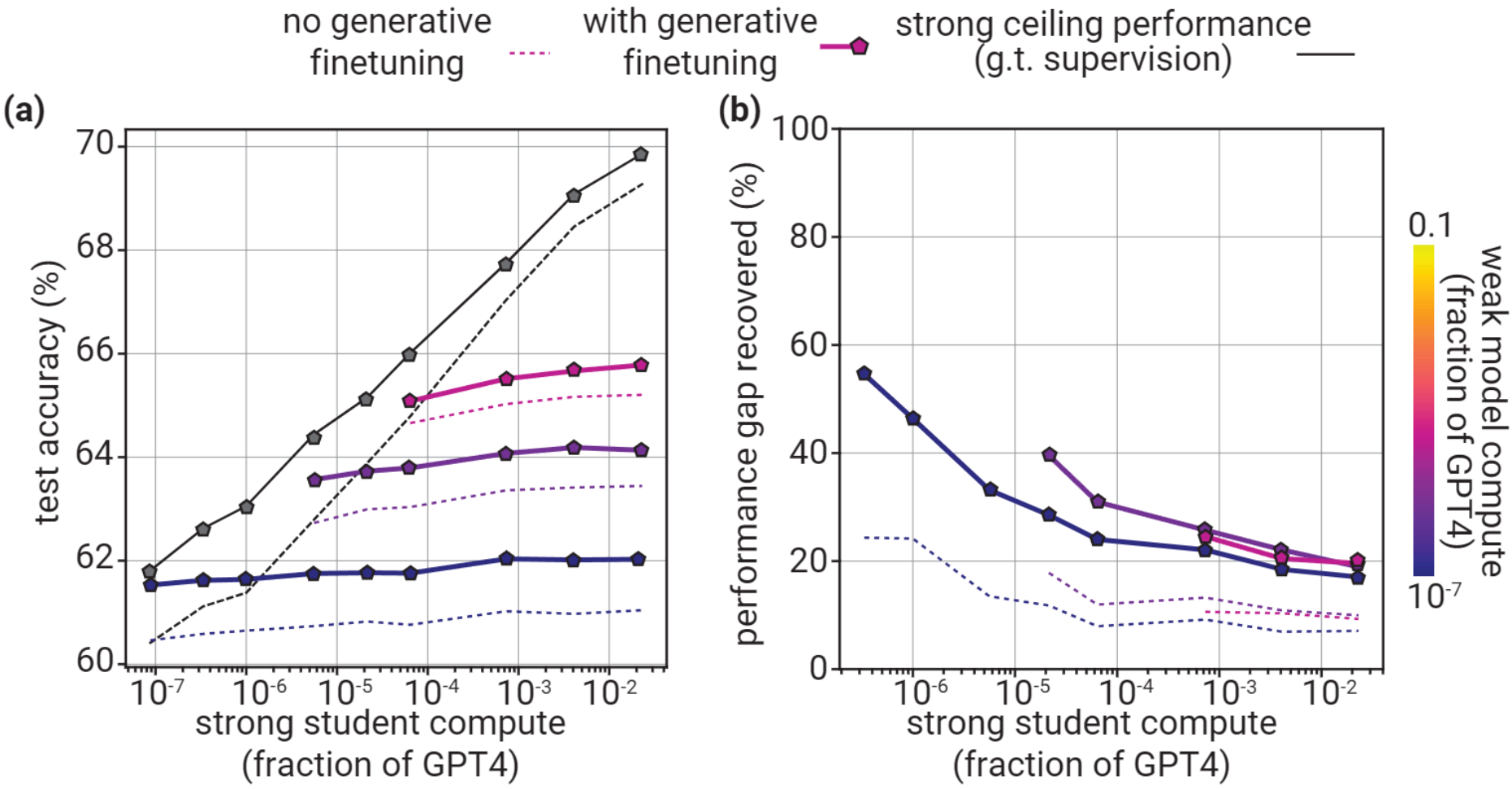
- **Idea**: Training models with a sequence of increasing model sizes, $\mathcal{M}_1 \rightarrow \mathcal{M}_2 \cdots \rightarrow \mathcal{M}_n$
- Test accuracy improves for larger student models in **Chess** task.



Improving Weak-to-Strong Generalization

Unsupervised Finetuning

- **Idea**: Unsupervised finetuning to increase salience of a task without ground truth labels
- PGR improves 10-20% in **ChatGPT RM** task.



Additional Analysis: Imitation

Q. What extent does the student imitate the weak supervision?

1. **Overfitting to weak model errors**
   - The student model overfits to weak model errors in early stage. (less than 1 epoch)
   - Oracle early stopping improves performance significantly.
   - Early stopping methods without relying on an oracle could enhance performance.
2. **Student-supervisor agreement**
   - Student-supervisor agreement is higher than weak supervisor accuracy.
   - Student imitates some of the supervisor's errors.
   - Confidence loss reduces this agreement.
3. **Size of student model**
   - Larger student models agree less with the supervisor than smaller models.
   - Contradictorily, larger models should fit teacher errors more easily.

Additional Analysis: Salience

Intuition: A capability should be easier to elicit is a model is salient.

1. **(Few-shot) Prompting vs. Finetuning**
   - Few-shot prompting becomes competitive with finetuning at large model sizes.
   - Weak-to-strong finetuning with confidence loss outperforms few-shot prompting.
   - Weak-to-strong fintuning significantly better than prompting at small model sizes.
2. **Linear Representation of Tasks**
   - Finetuning on weak labels makes the ground truth task more linearly represented.
   - Ground truth linear probe of the base model: 72%
   - Finetuning on ground truth: 82%
   - Ground truth linear probe of a model finetuned on weak labels: 78%.

Remaining Disanalogies

1. **Imitation saliency: superhuman models may easily imitate weak errors.**
   - If we train models with human supervision, it might be having human-level capabilities.
   - Larger strong models have more difficulty to imitate weak model.
   - **Imitationg the weak supervisor** may not be as much of a problem.

2. **Pretraining leakage: superhuman knoledge may be latent, not observable.**
   - Many of tasks may have been observed in pretraining stage. (contamination)
   - Future superhuman model may never directly observe alignment-relevant capabilities.