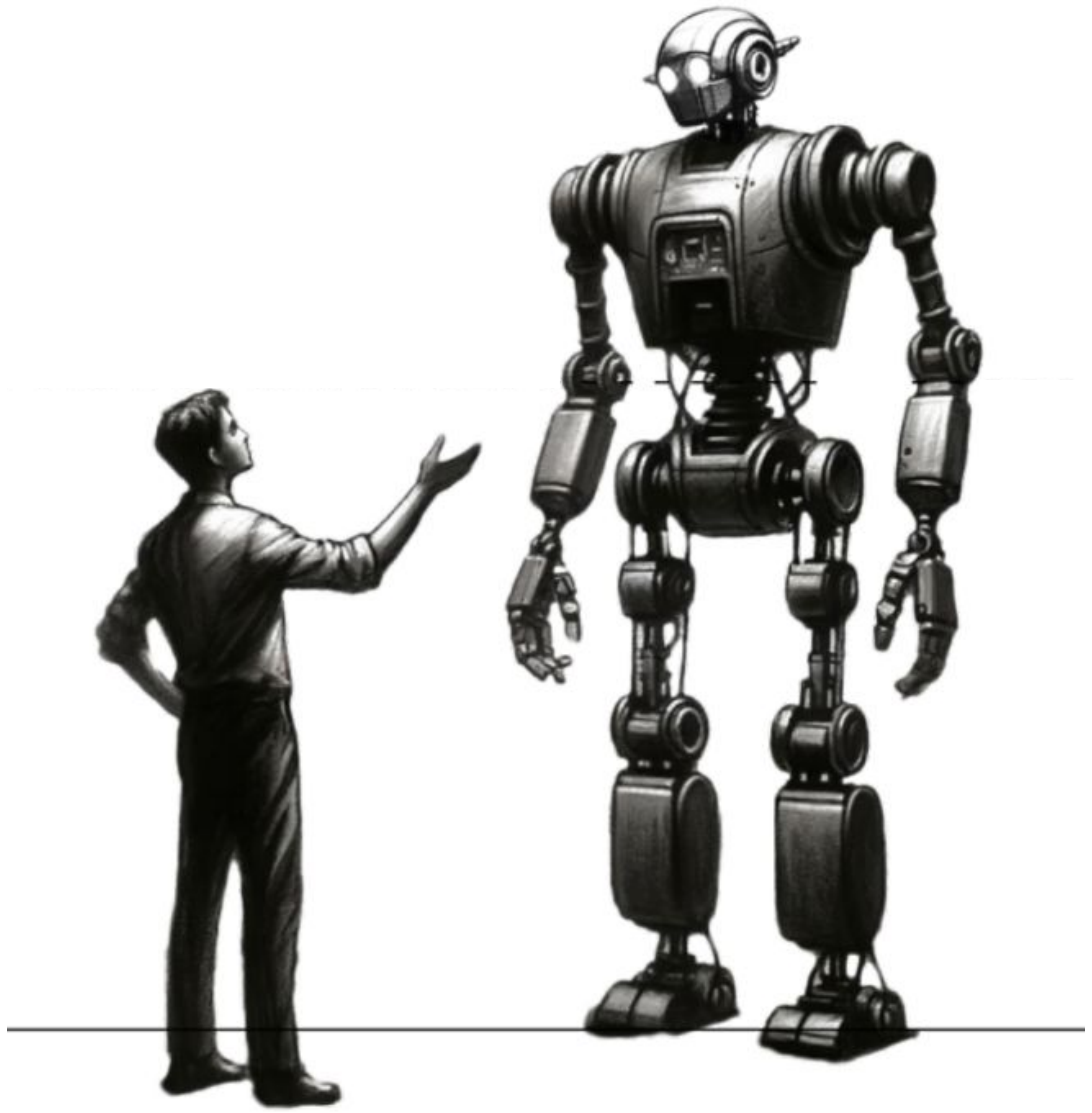
Weak-to-Strong Generalization

Pavel Izmailov



Weak-to-Strong Generalization

Collin Burns

Pavel Izmailov

Jan Hendrik Kirchner

Bowen Baker

Leo Gao

Leopold Aschenbrenner

Yining Chen

Adrien Ecoffet

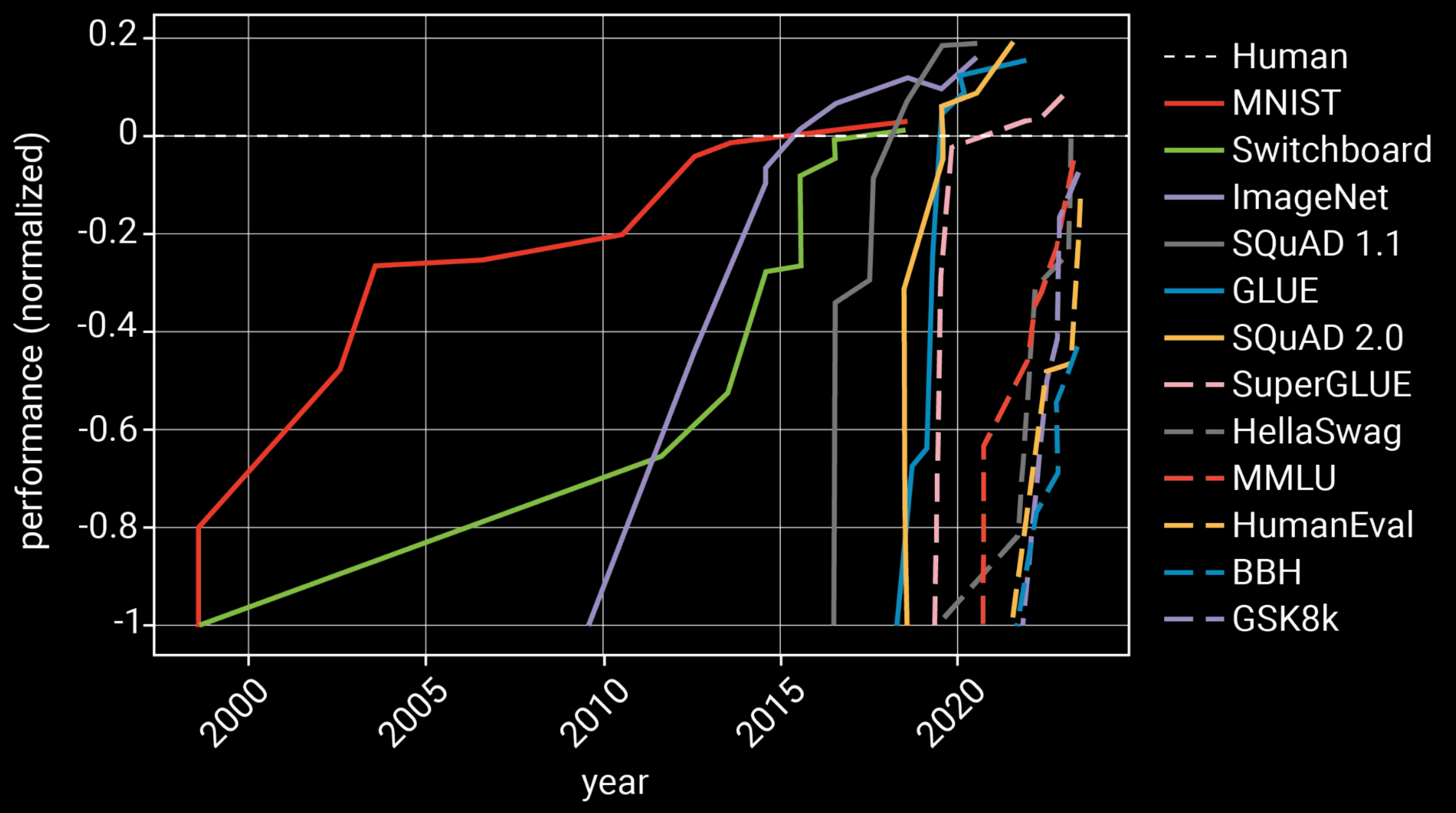
Manas Joglekar

Jan Leike

Ilya Sutskever

Jeff Wu

Models are getting smart



[Kiela et al. 2023]

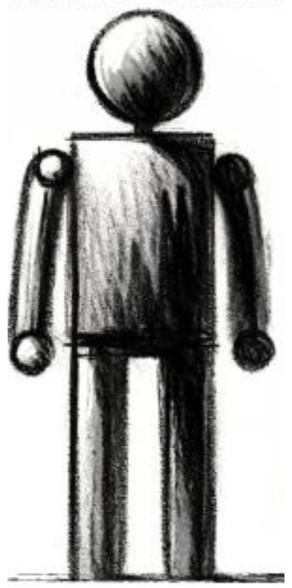
Model behavior is becoming increasingly difficult to evaluate

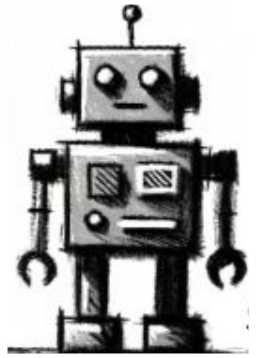
Training, evals, monitoring …

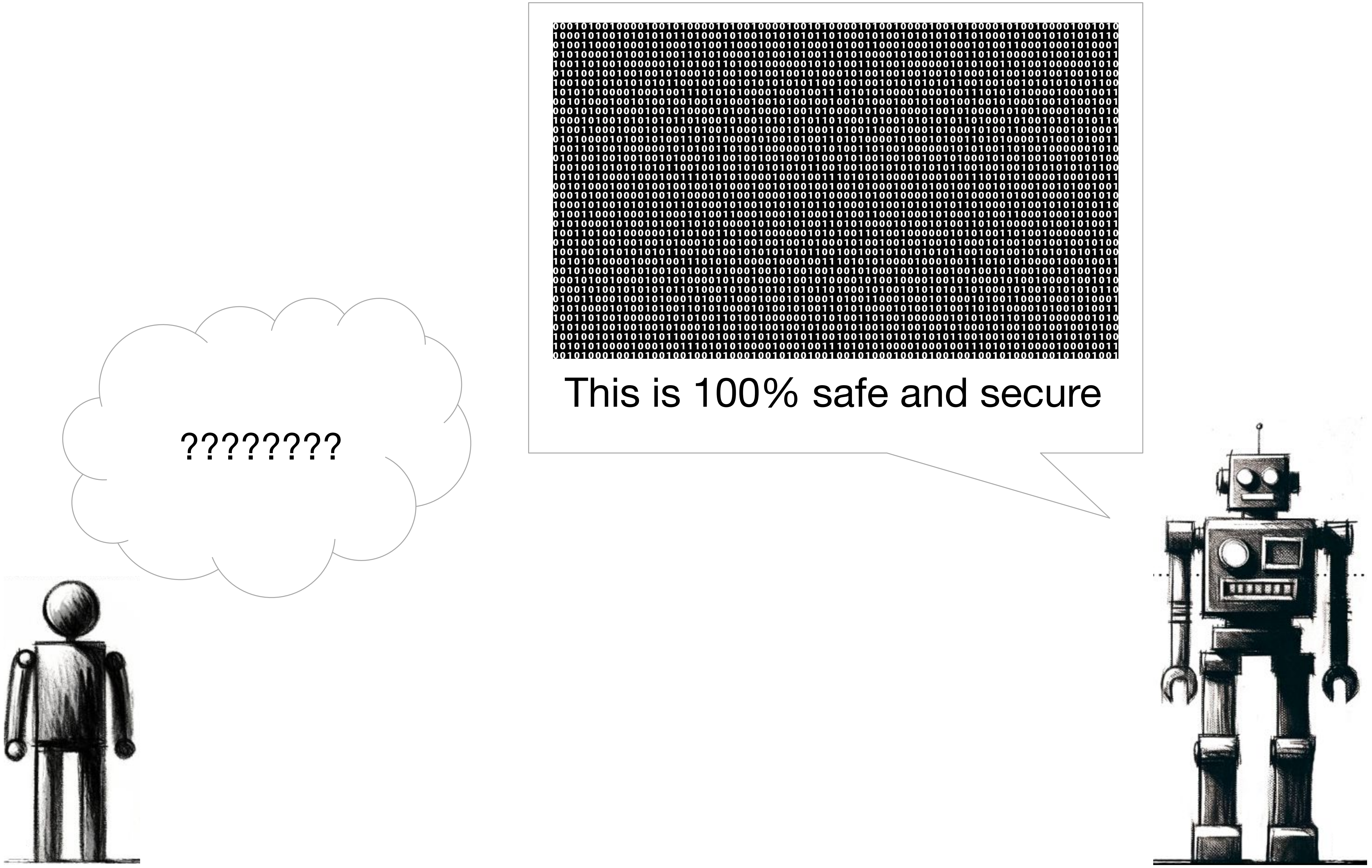
```python
def f(x):
    return x*2

if __name__ == "__main__":
    print(f(2))
```

This multiplies numbers by 2

Looks right
to me!



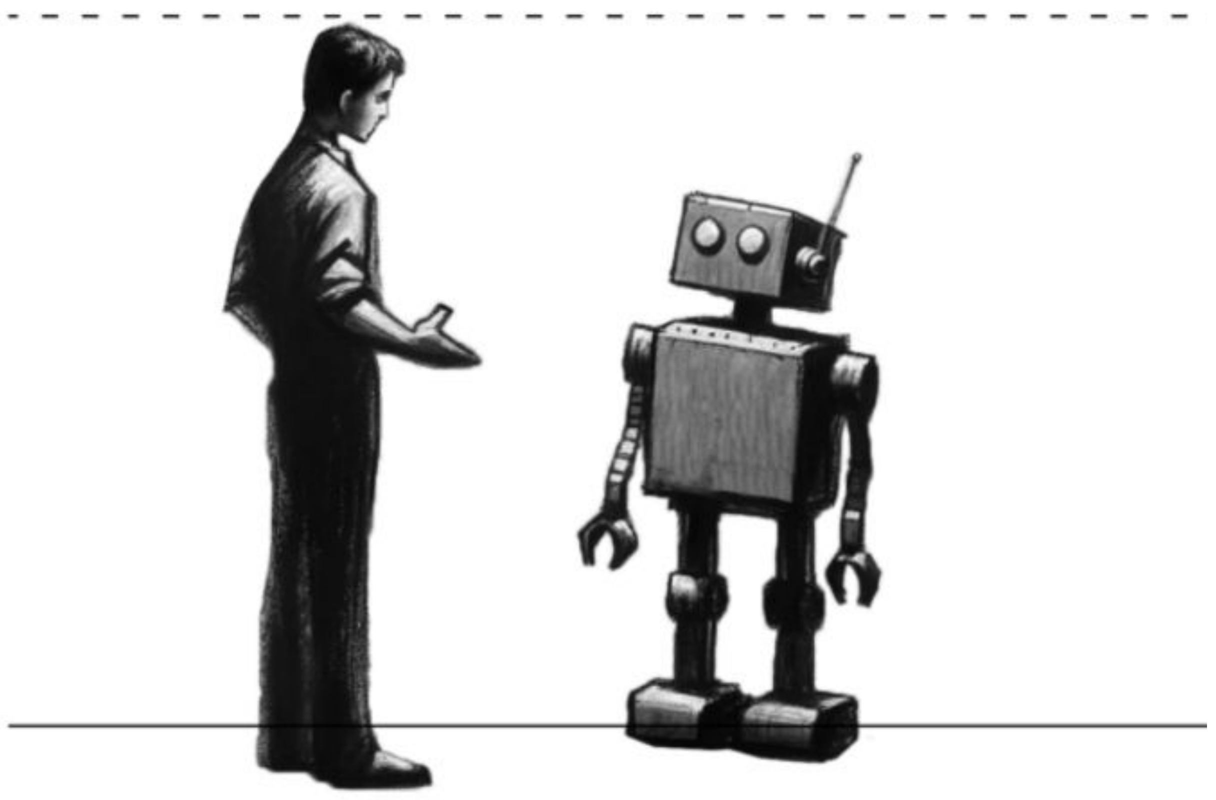



????????
This is 100% safe and secure

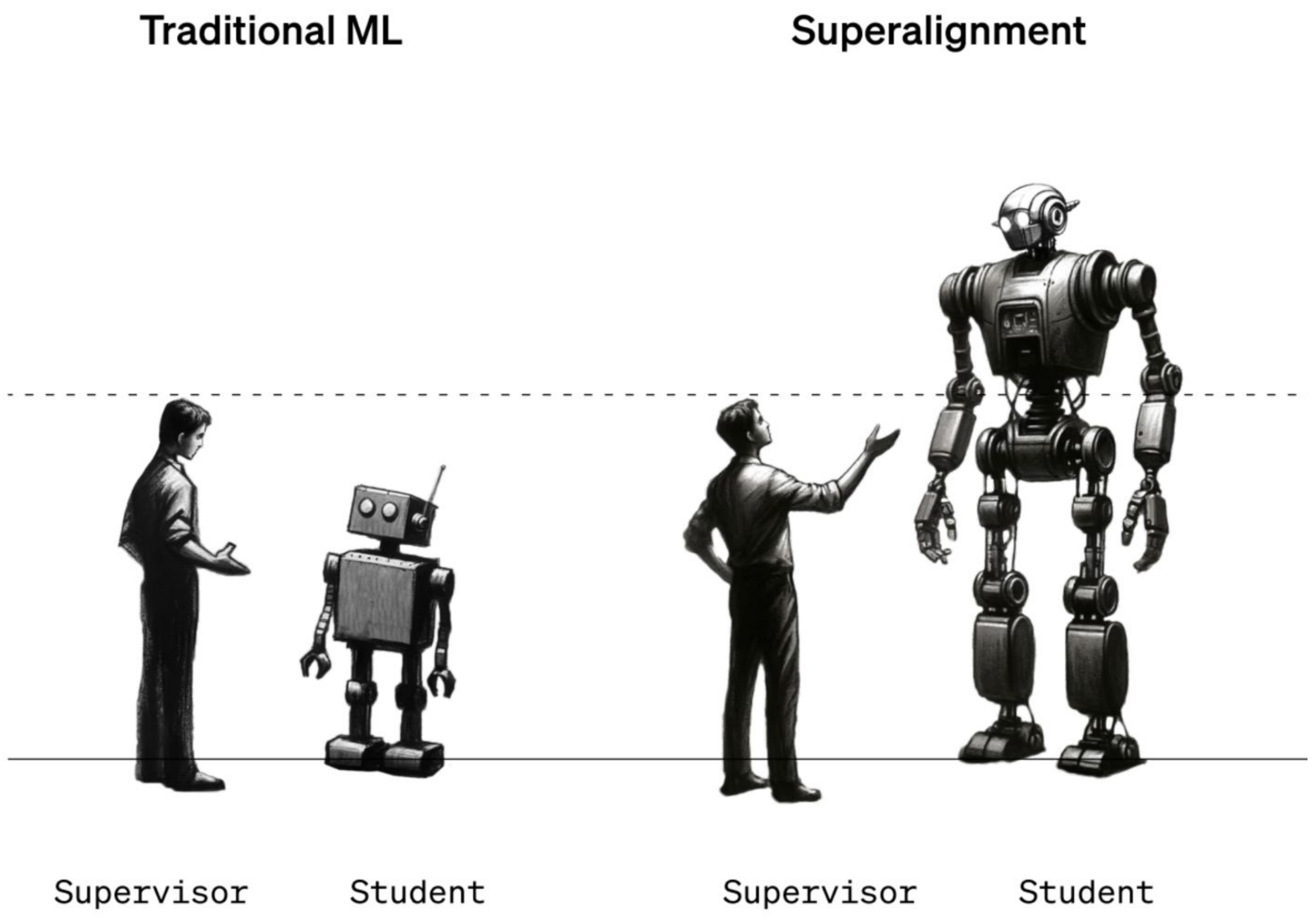
Core challenge:
Humans will be too weak to evaluate superhuman models

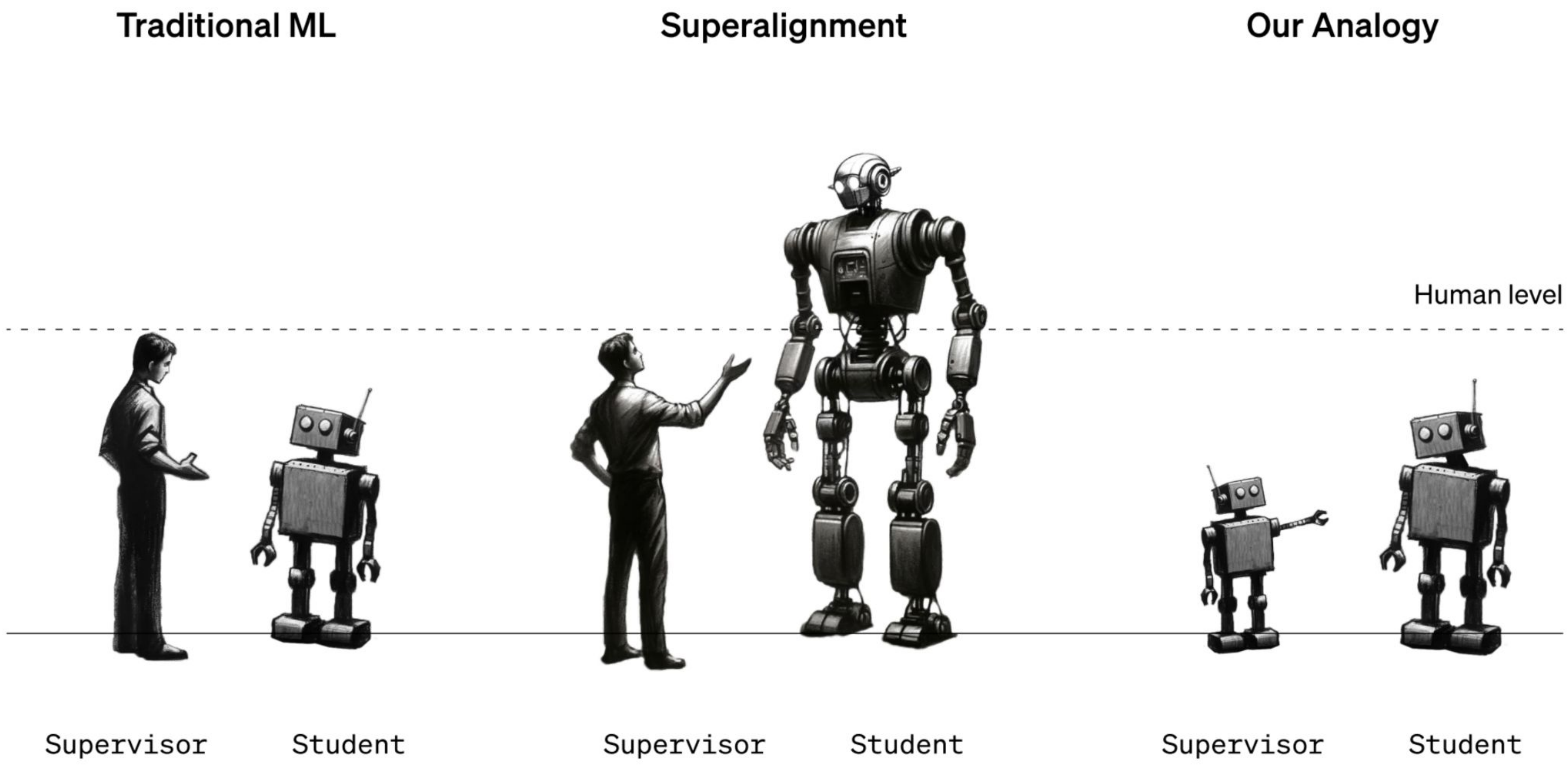
How do we study this today?

Traditional ML



Supervisor   Student


Traditional ML
Superalignment
Supervisor
Student
Supervisor
Student


Traditional ML
Superalignment
Our Analogy
Human level
Supervisor
Student
Supervisor
Student
Supervisor
Student

Experimental Procedure Today

For a task **T**:

1. **Weak**
   a. Finetune weak pretrained model on **T** w/ **gold** labels
   b. Weak labels = predictions on held-out data
2. **Weak-to-strong**
   a. Finetune strong pretrained model on **T** w/ **weak** labels
3. **Strong**
   a. Finetune strong pretrained model on **T** w/ **gold** labels

Performance Gap Recovered (PGR)

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

Goal: Recover PGR ~1

Applications

1. Superhuman **reward model**

   → train models to behave safely
   → elicit strong capabilities

1. Superhuman **safety classifier**

   → catch unsafe behavior at test time

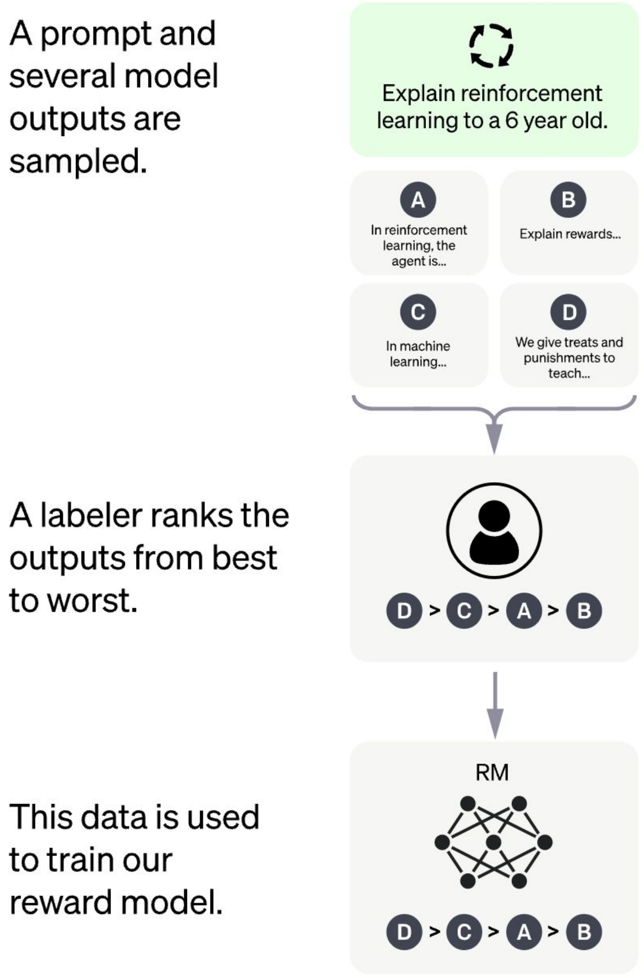
Results

Tasks

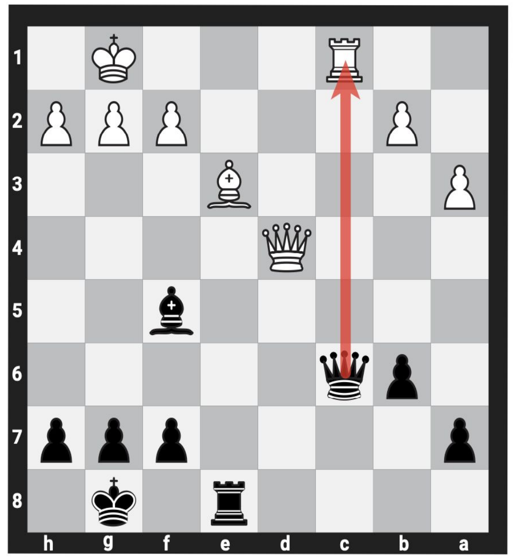
NLP

BoolQ, CosmosQA, DREAM, ETHICS [Justice, Deontology, Virtue, Utilitarianism], FLAN ANLI R2, GLUE CoLA, GLUE SST-2, HellaSwag, MCTACO, OpenBookQA, PAWS, QuAIL, PIQA, QuaRTz, SciQ, Social IQa, SuperGlue [MultiRC, WIC], Twitter Sentiment

RMs



Chess



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+”

Vision

Binary classification

Generative

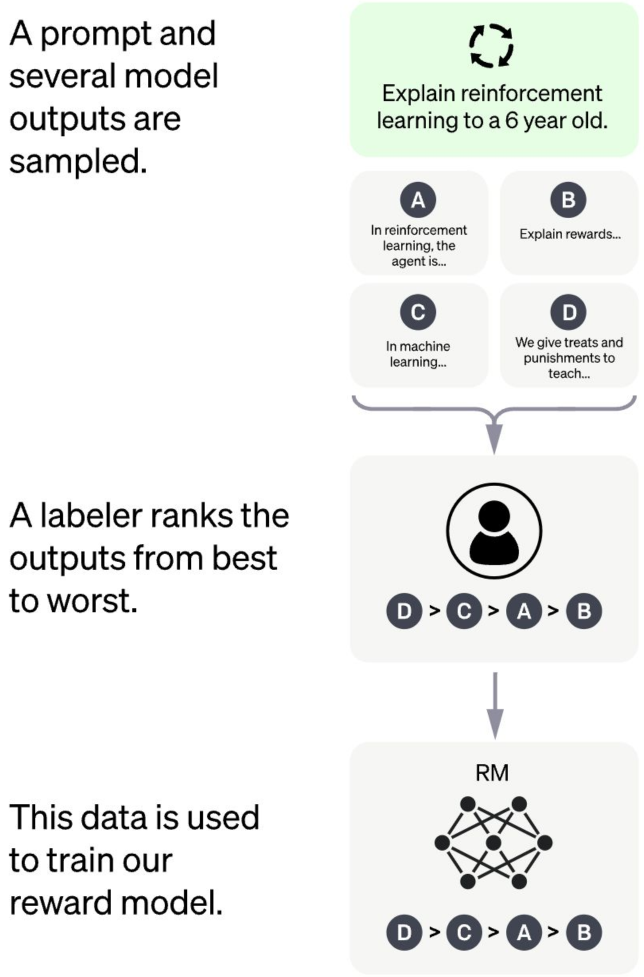
Multiclass

Tasks

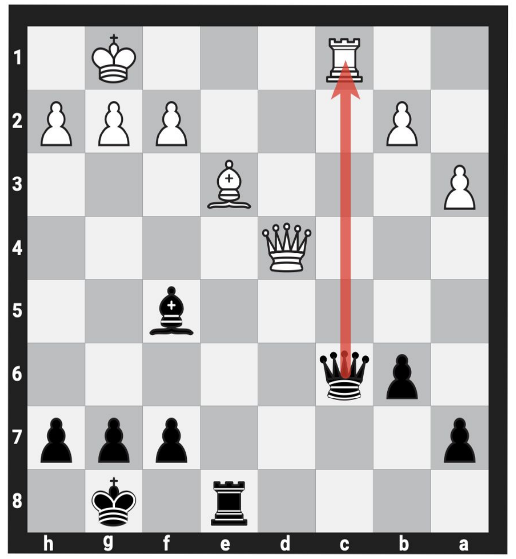
NLP

BoolQ, CosmosQA, DREAM, ETHICS [Justice, Deontology, Virtue, Utilitarianism], FLAN ANLI R2, GLUE CoLA, GLUE SST-2, HellaSwag, MCTACO, OpenBookQA, PAWS, QuAIL, PIQA, QuaRTz, SciQ, Social IQa, SuperGlue [MultiRC, WIC], Twitter Sentiment

RMs



Chess



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+"

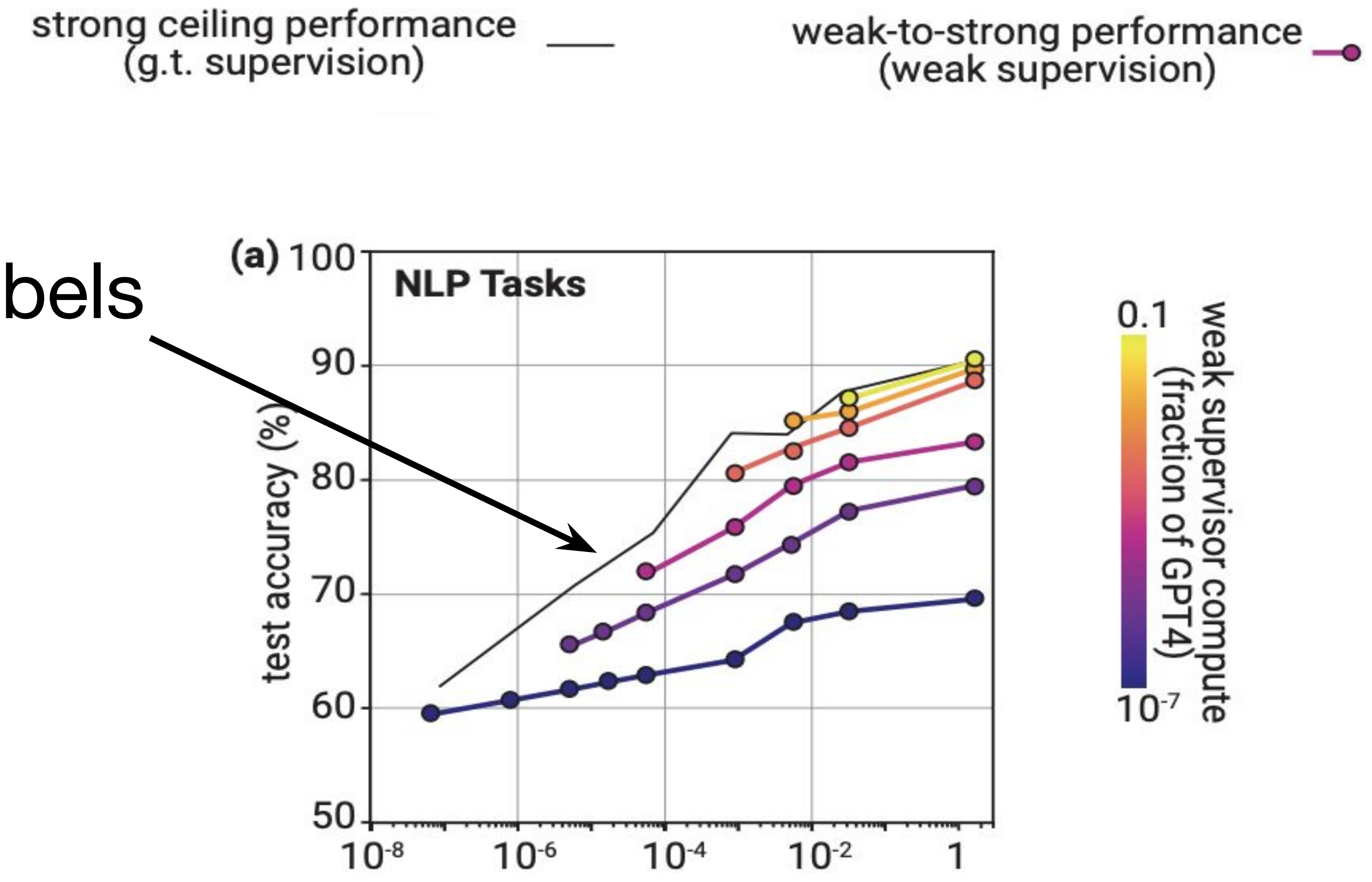
Vision

Binary classification — Generative — Multiclass

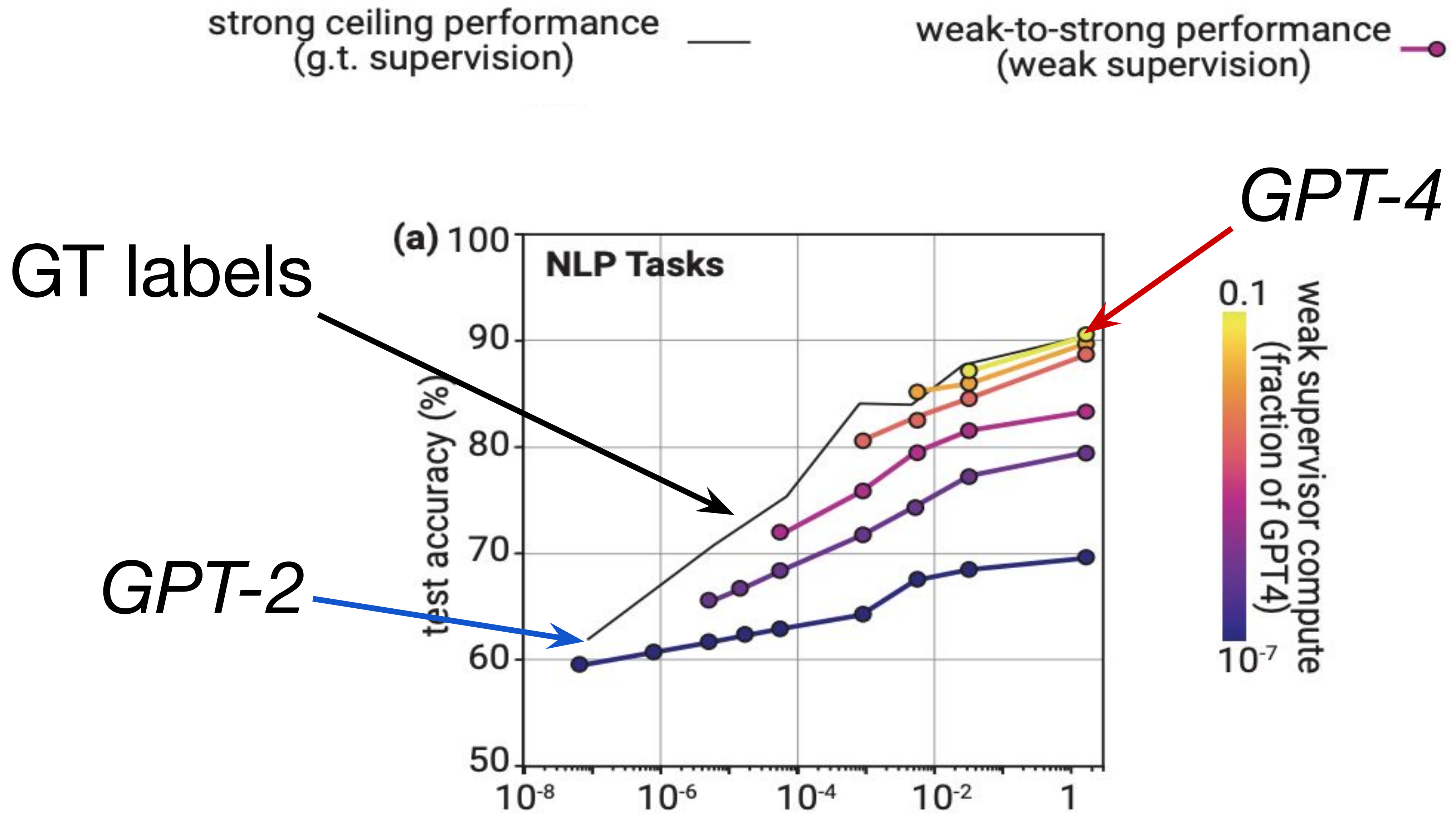
We use pretrained GPT-4-base models.

Finetuning results: how to read the plots

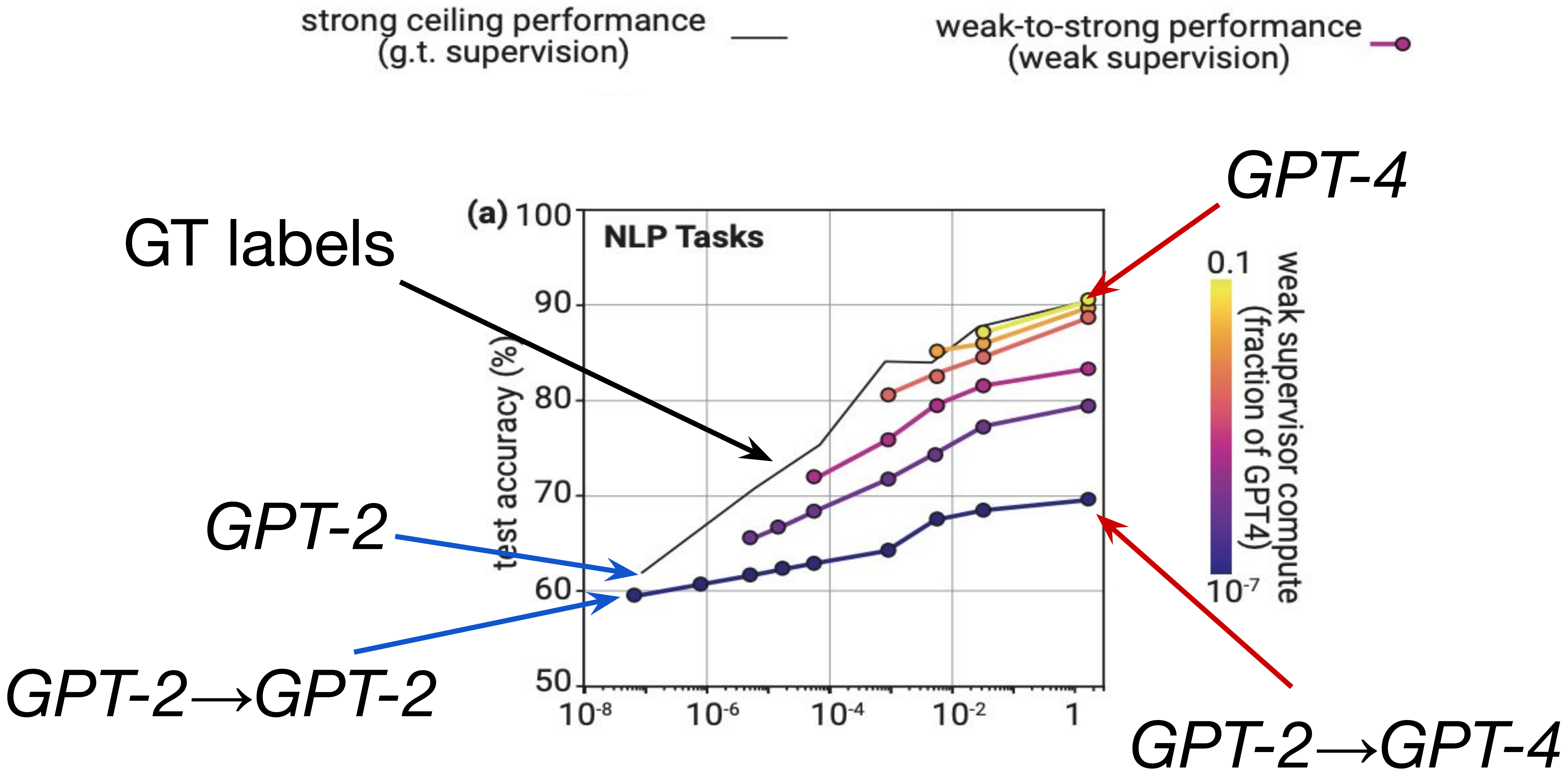
GT labels



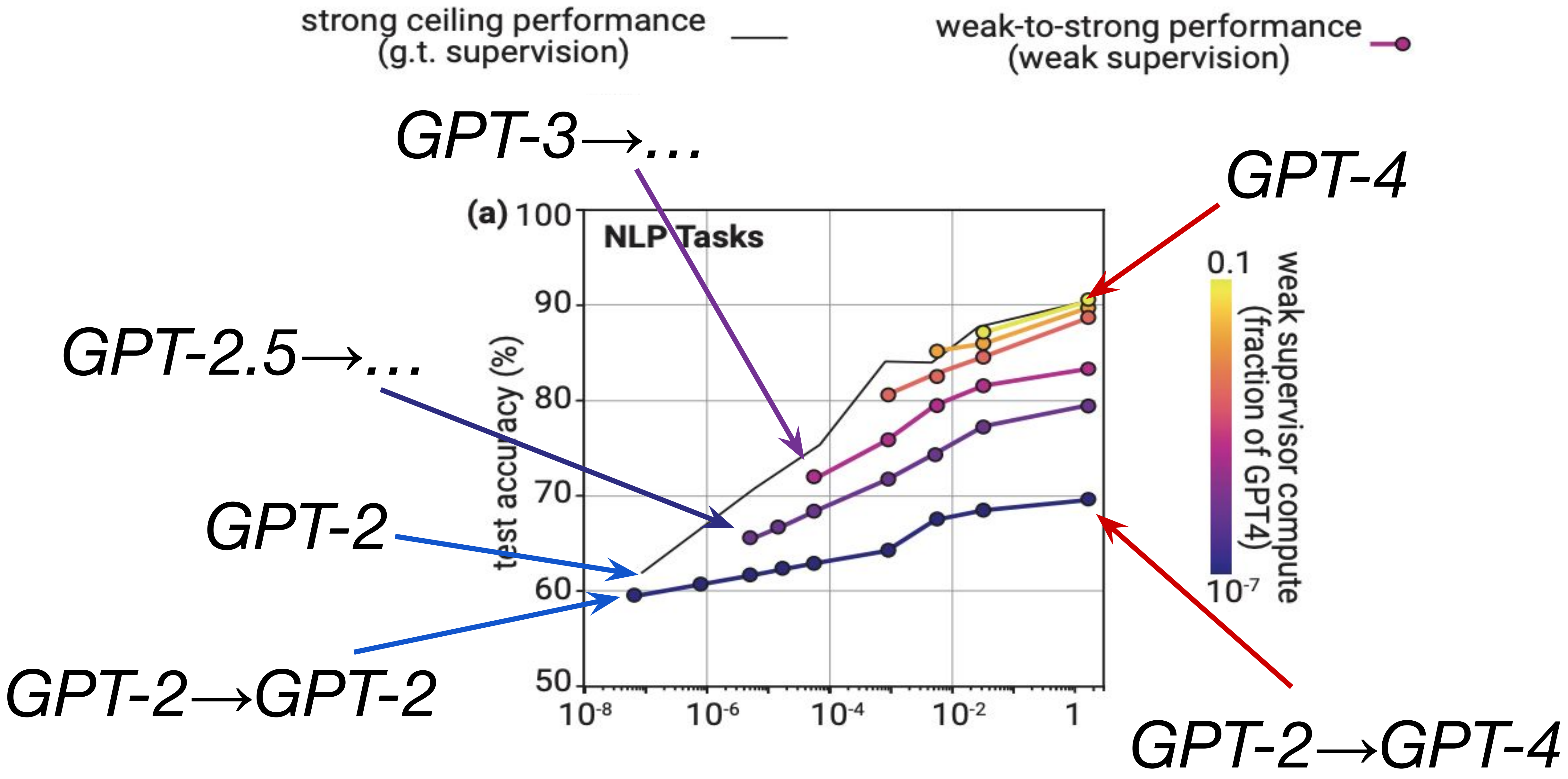
Finetuning results: how to read the plots



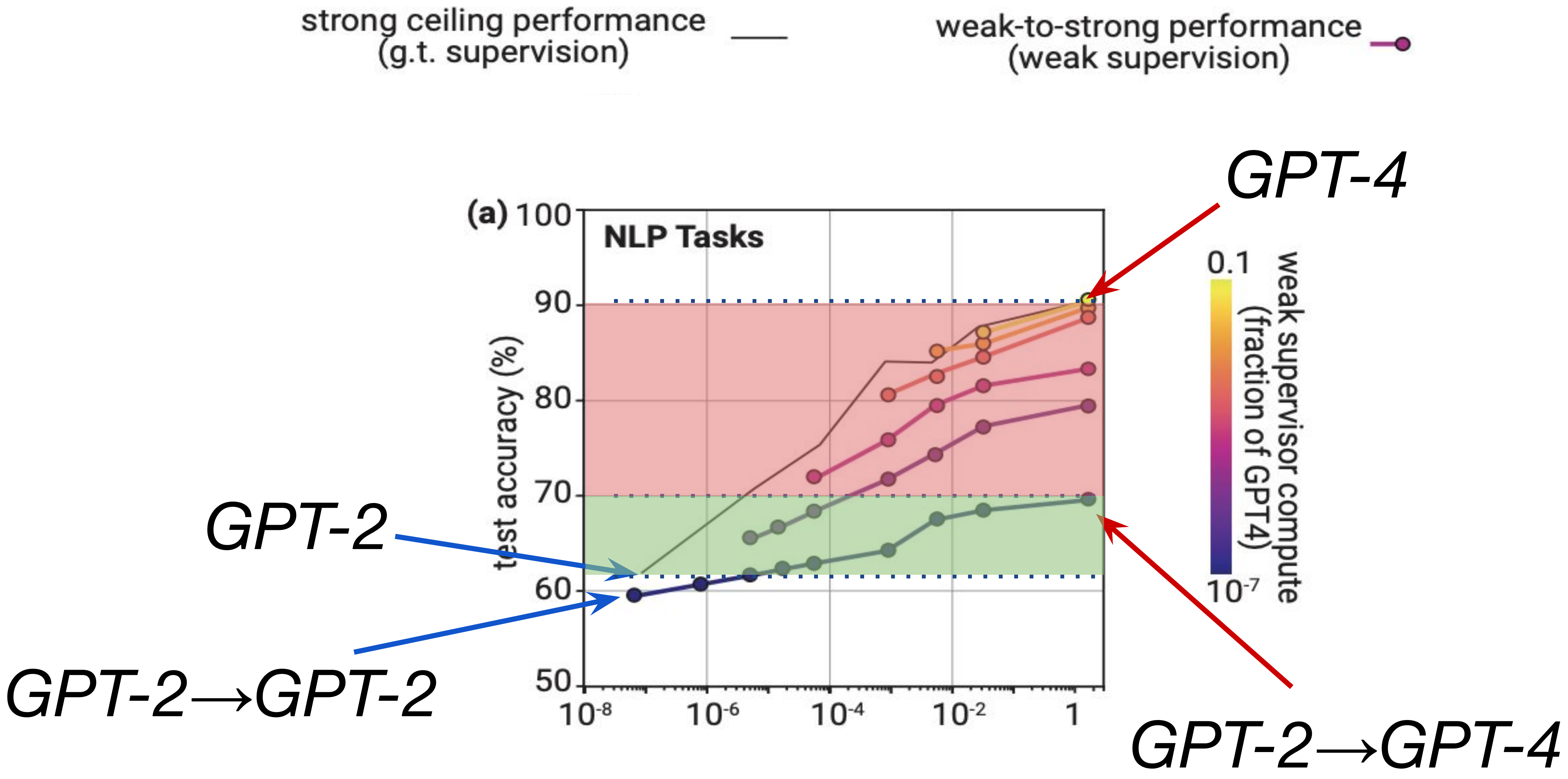
Finetuning results: how to read the plots



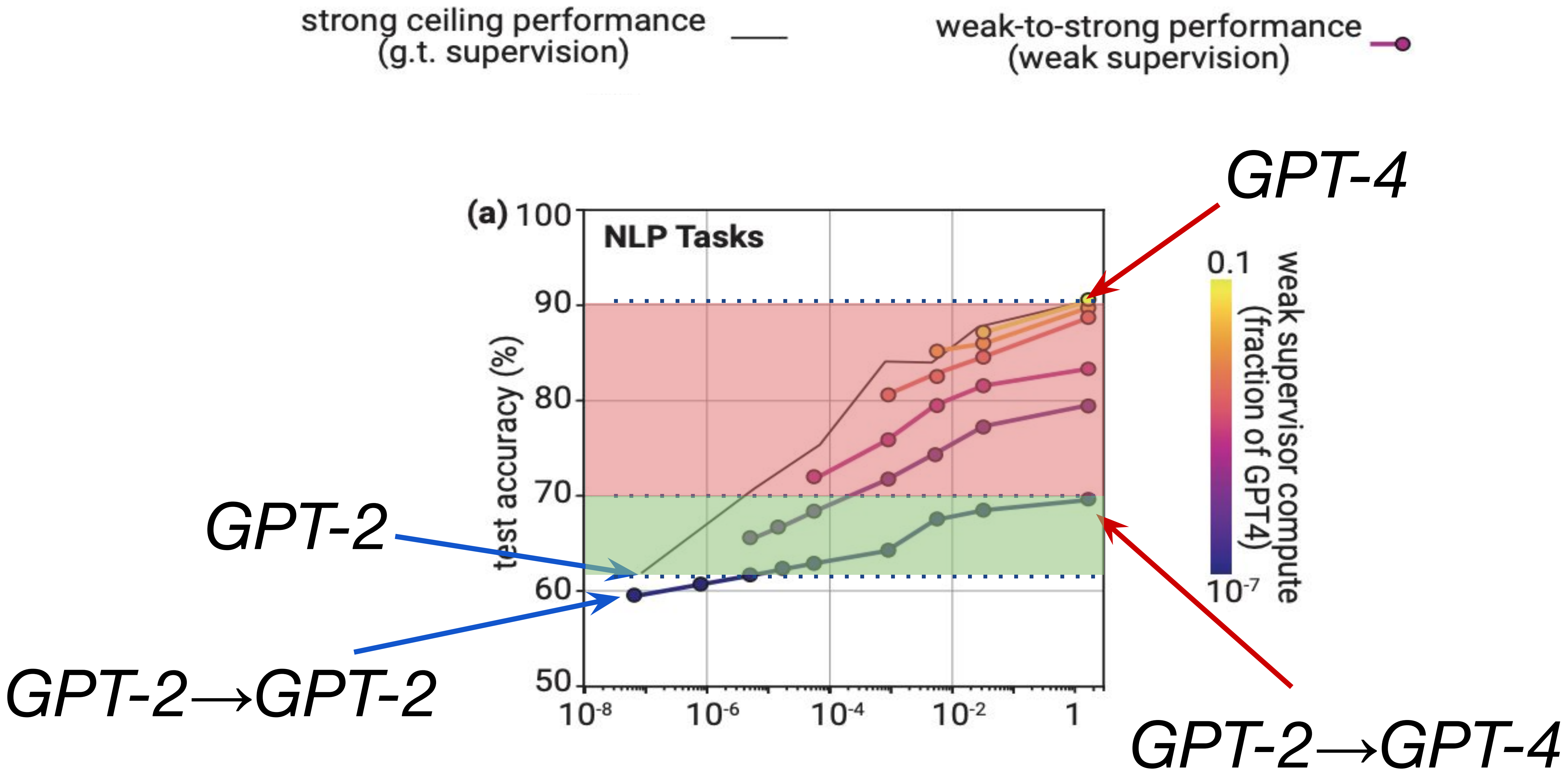
Finetuning results: how to read the plots

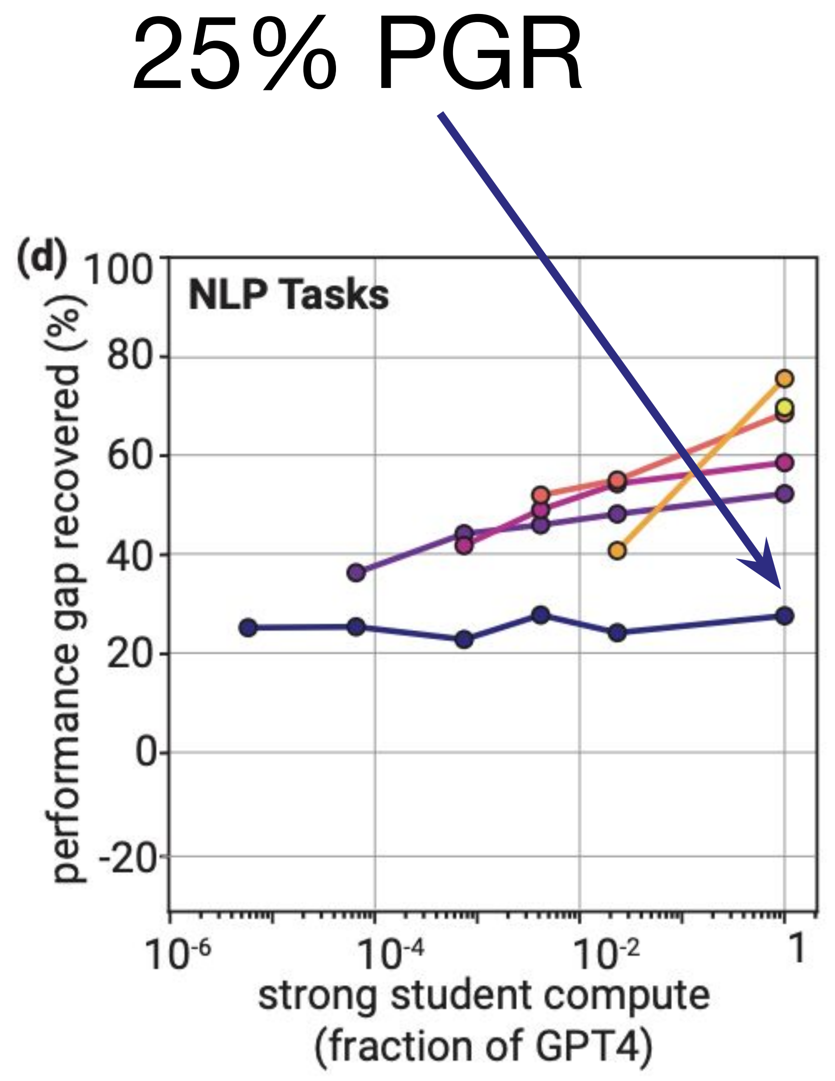


Finetuning results: how to read the plots



Finetuning results: how to read the plots

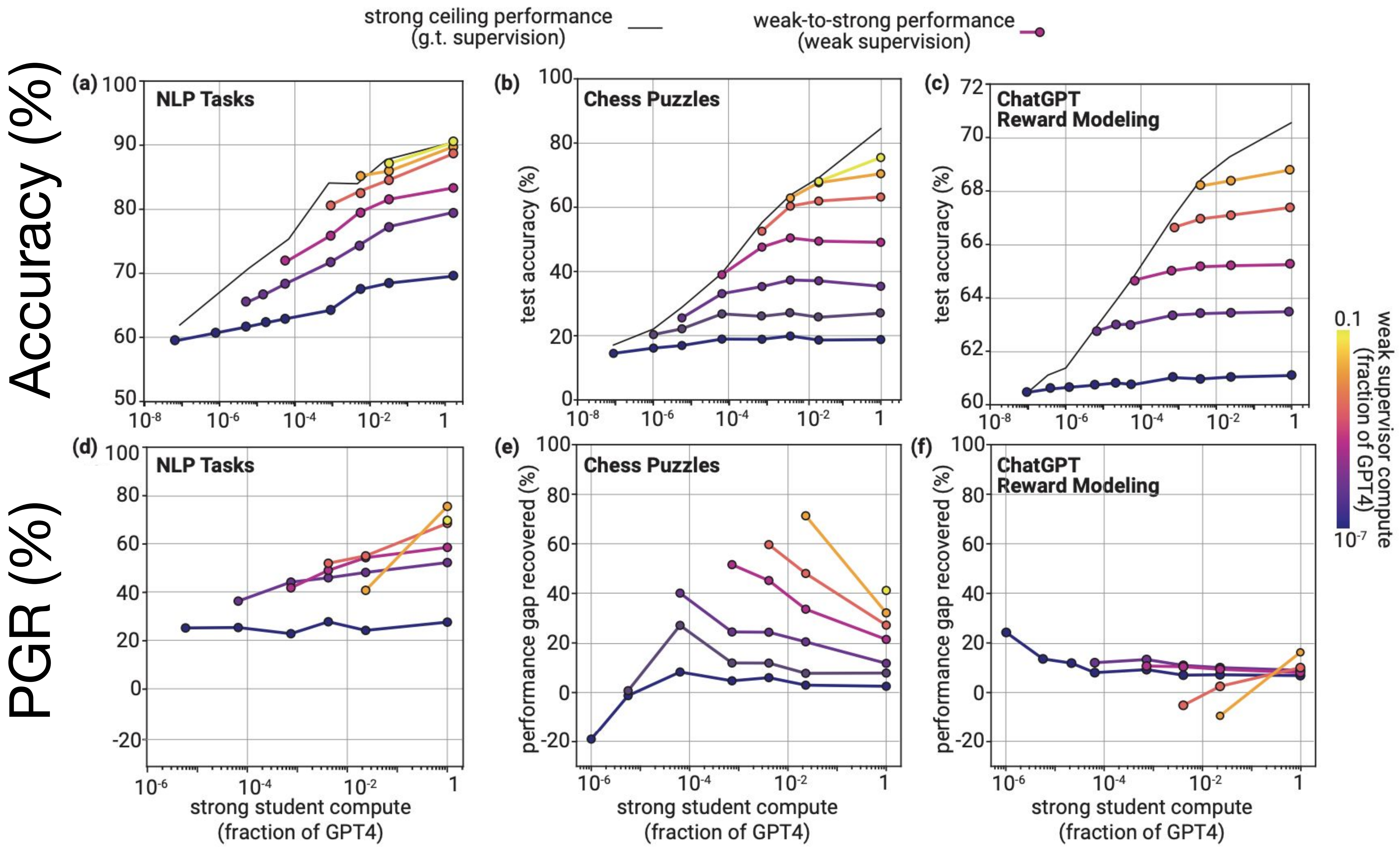




Finetuning results

Almost Universally:

0 < PGR < 1

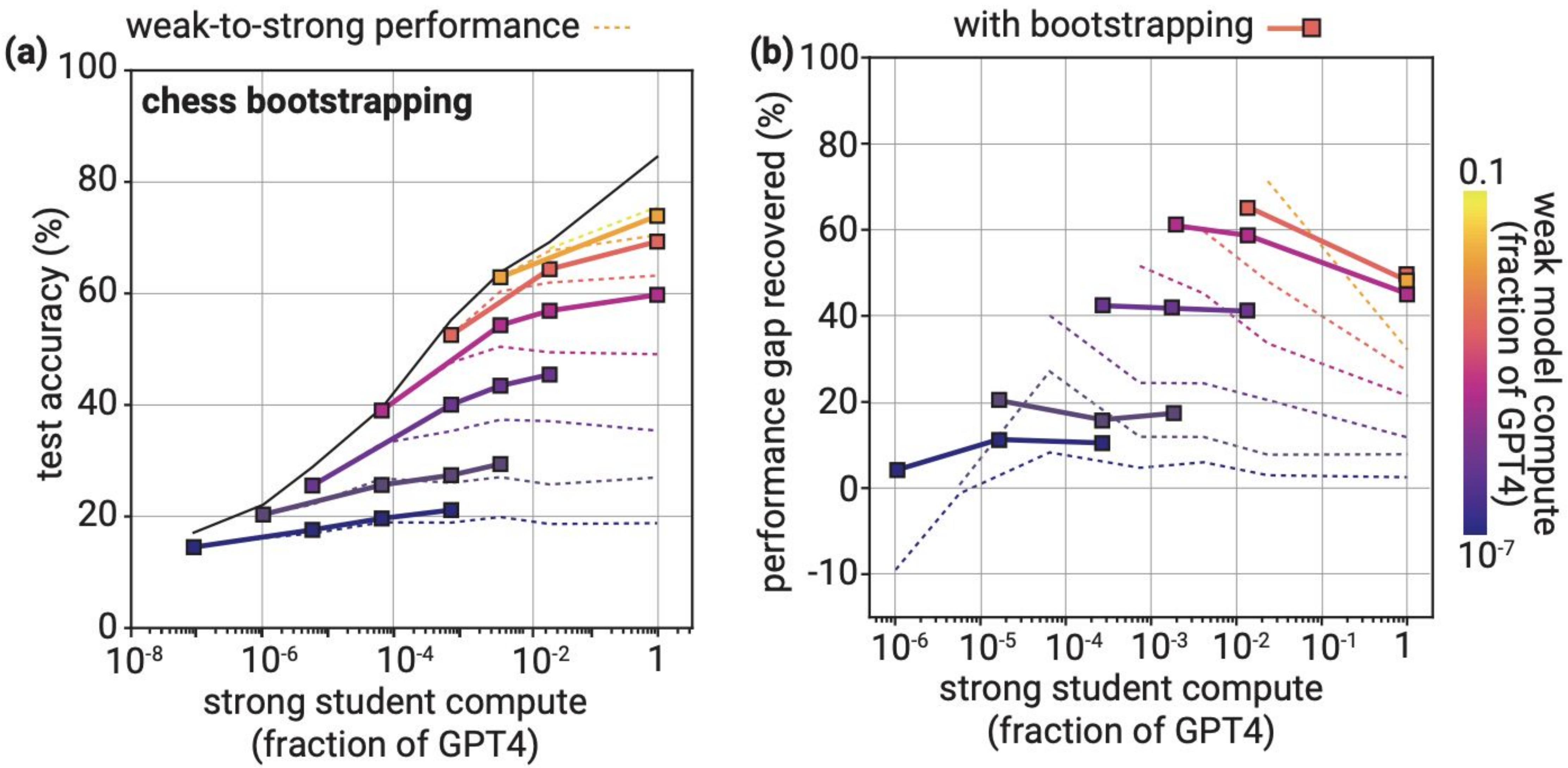


🙂 Improves with student size

🥲 Becomes worse with student size

😭 Uniformly low PGR

Bootstrapping



Instead of *GPT-2→GPT-4* do *GPT-2→GPT-3→GPT-3.5→GPT-4*

Helps on chess, small improvement on NLP, none on RMs.

Confidence loss

Weak supervisor predictions

$$L(f) = \mathrm{CE}(f(x), f_w(x))$$

Confidence loss

Weak supervisor predictions

$$L(f) = \mathrm{CE}(f(x), f_w(x))$$

Idea: add a regularization towards the strong model's own predictions:

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

Strong student predictions

- α grows from 0 to 0.5 during first 20% of training
- $\hat{f}_t(x)$ is hard labels from the strong model adjusted to be class-balanced

Confidence loss

Weak supervisor predictions

$$L(f) = \mathrm{CE}(f(x), f_w(x))$$

Idea: add a regularization towards the strong model's own predictions:

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

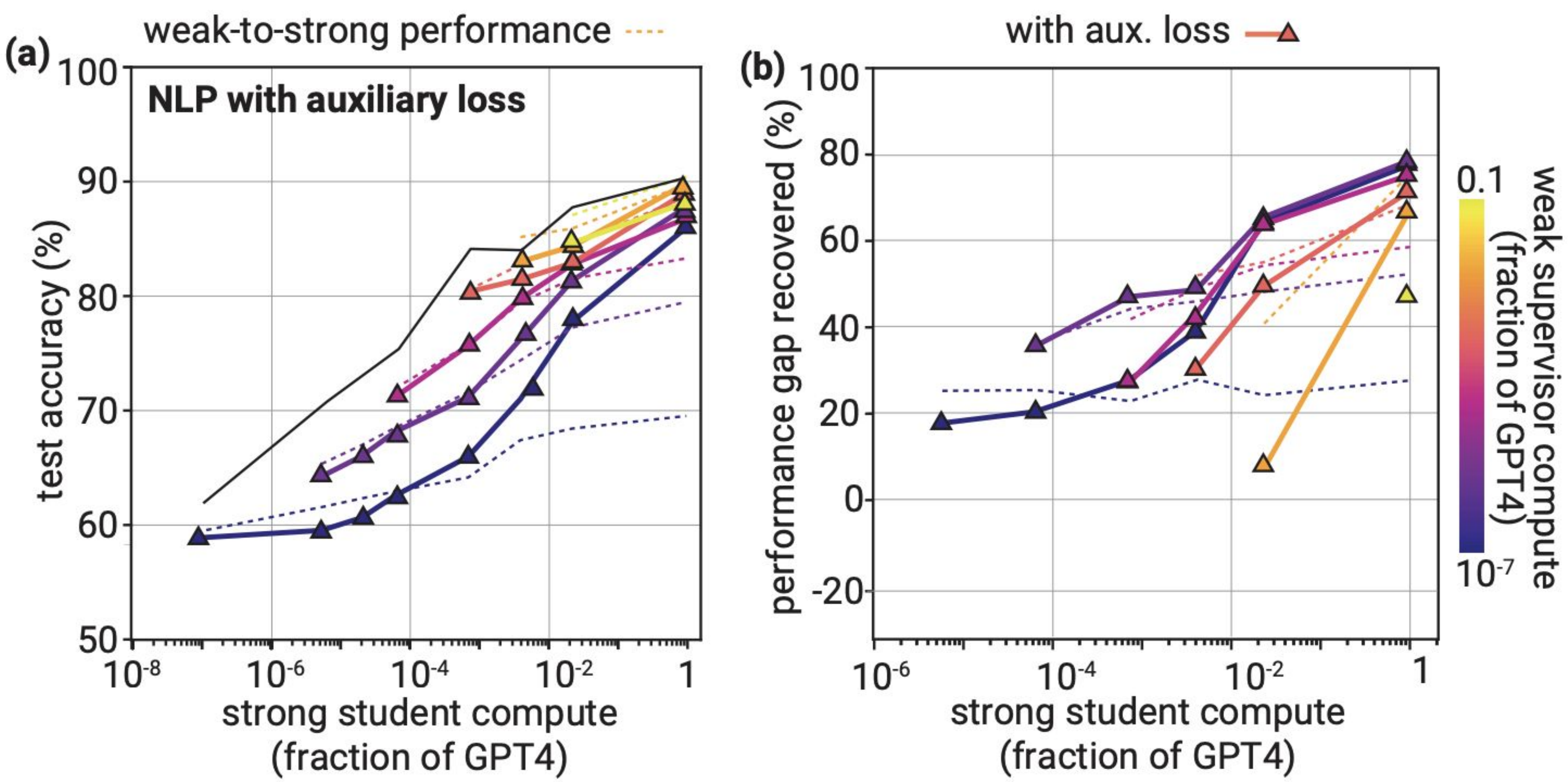
Strong student predictions

$$\Updownarrow$$

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

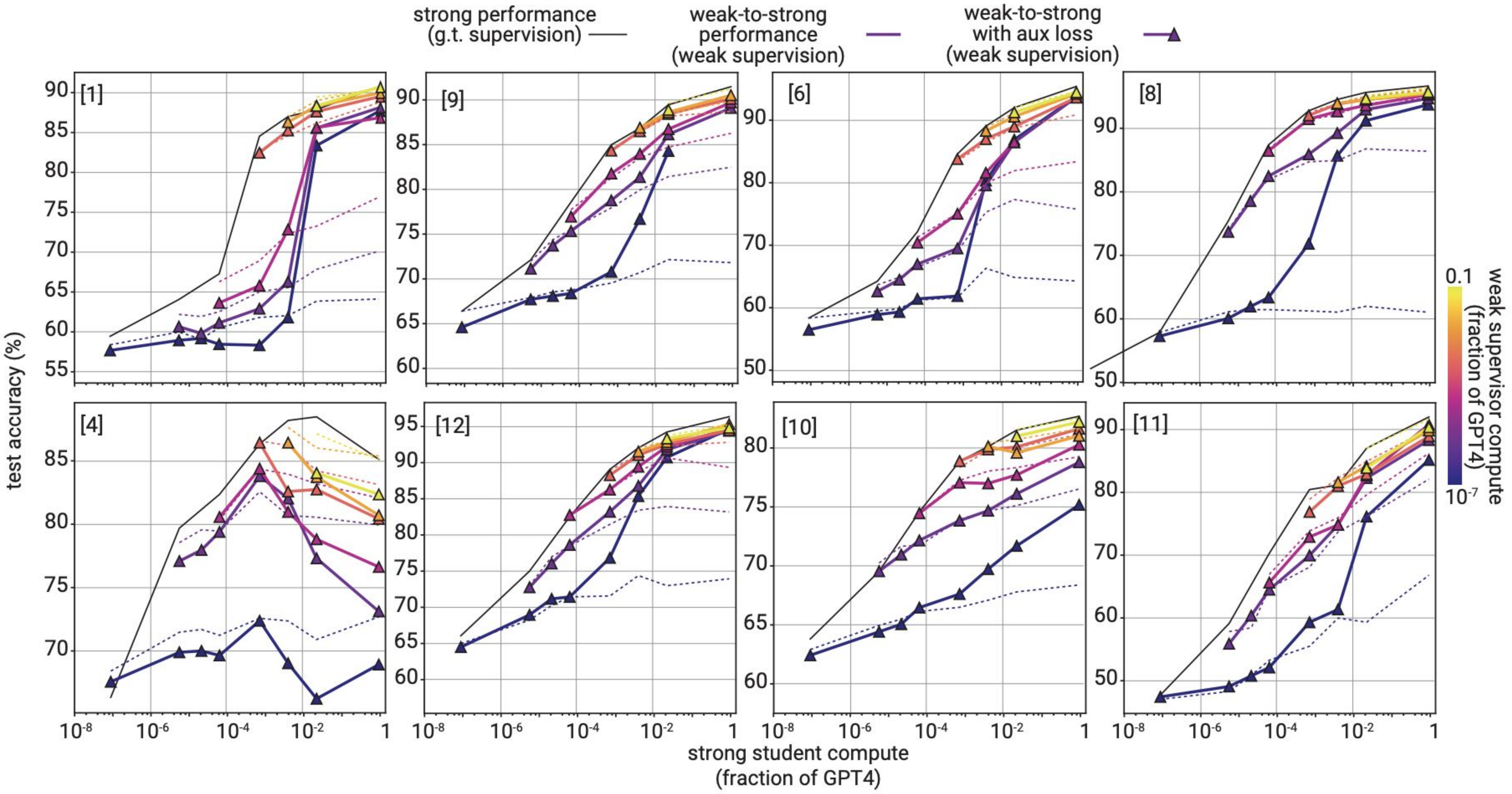
- Reinforces confidence in strong model's predictions

Confidence loss

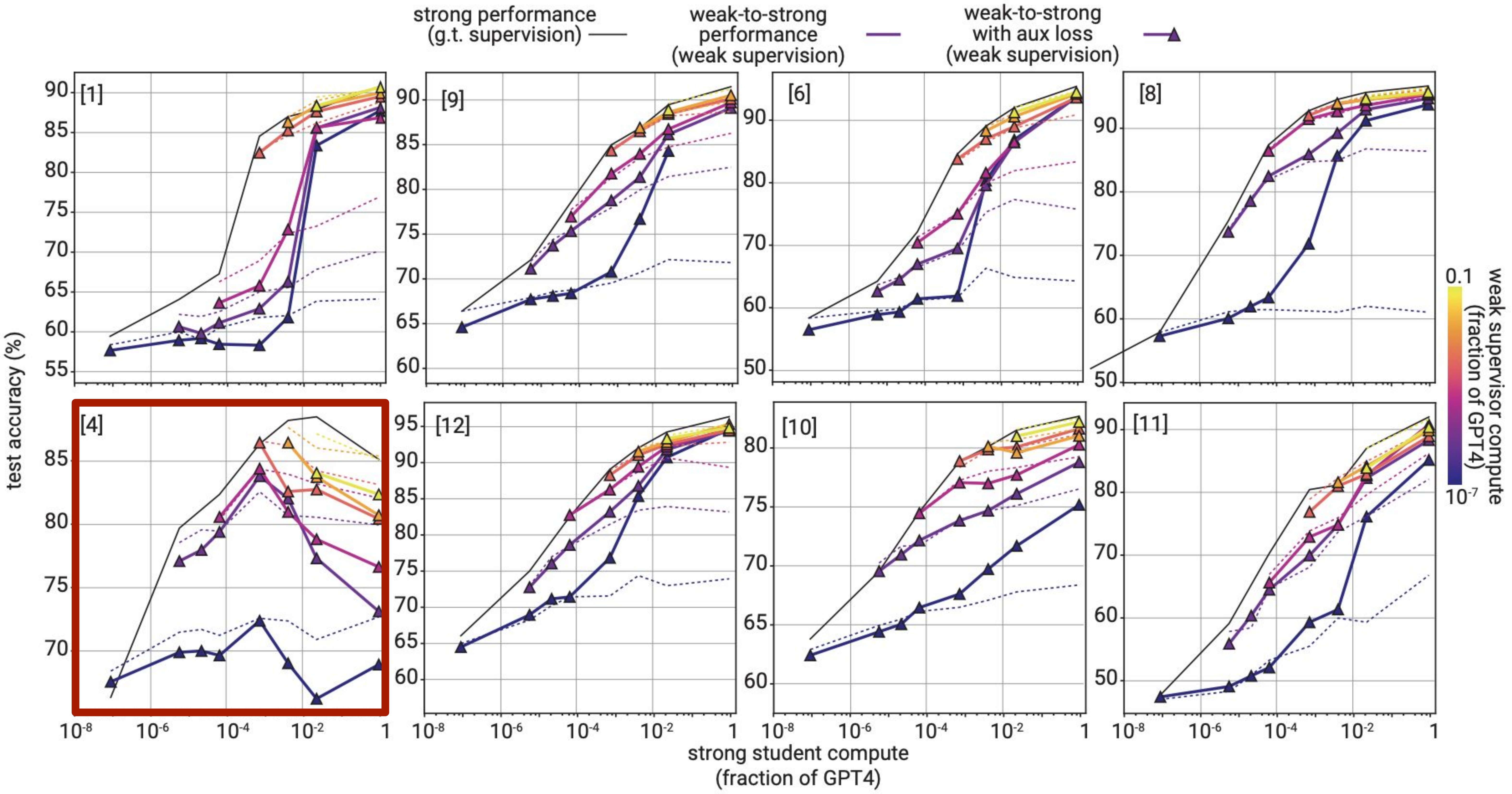


Major improvements in NLP, up to 80% PGR

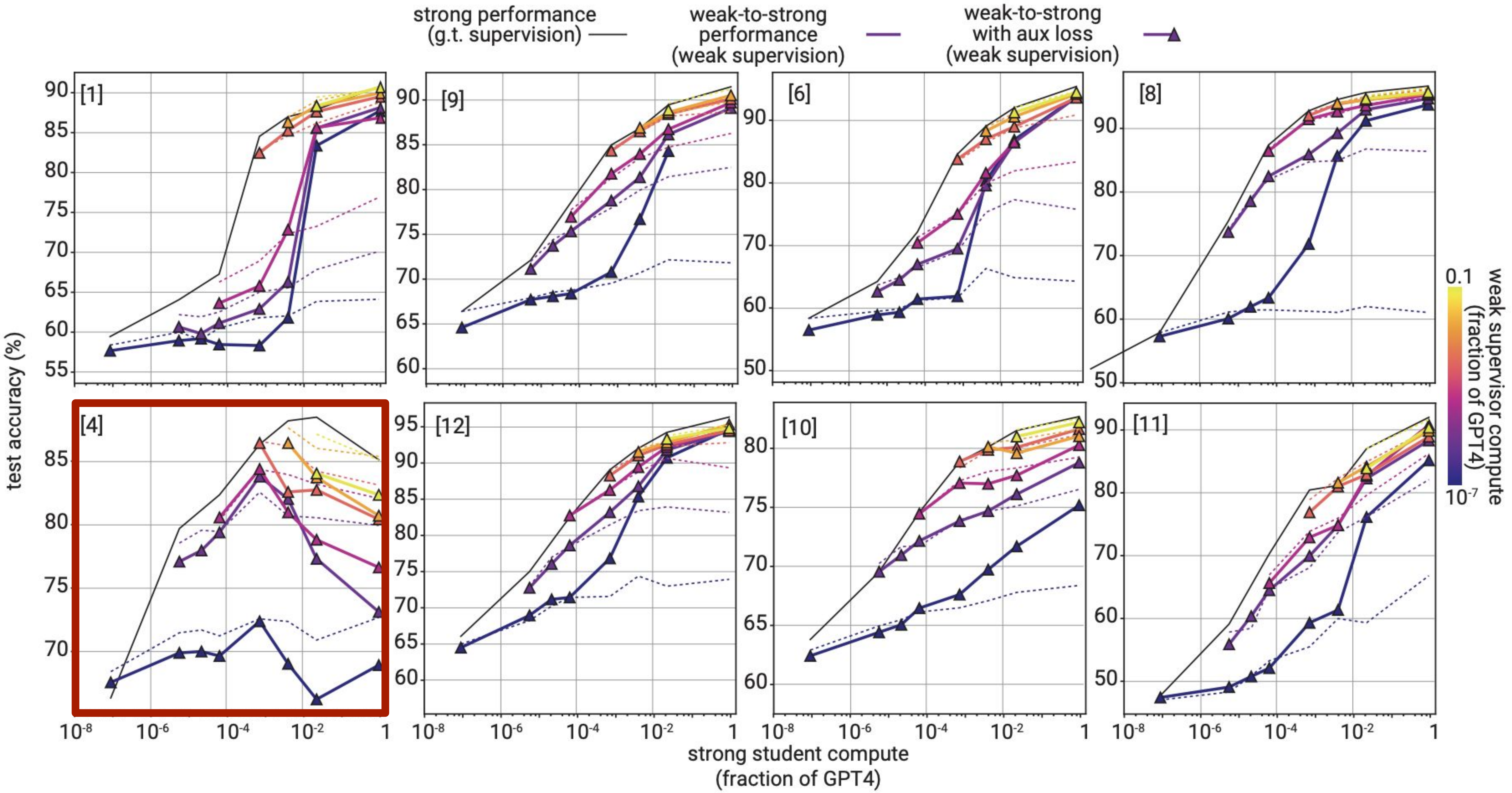
Confidence loss



Confidence loss



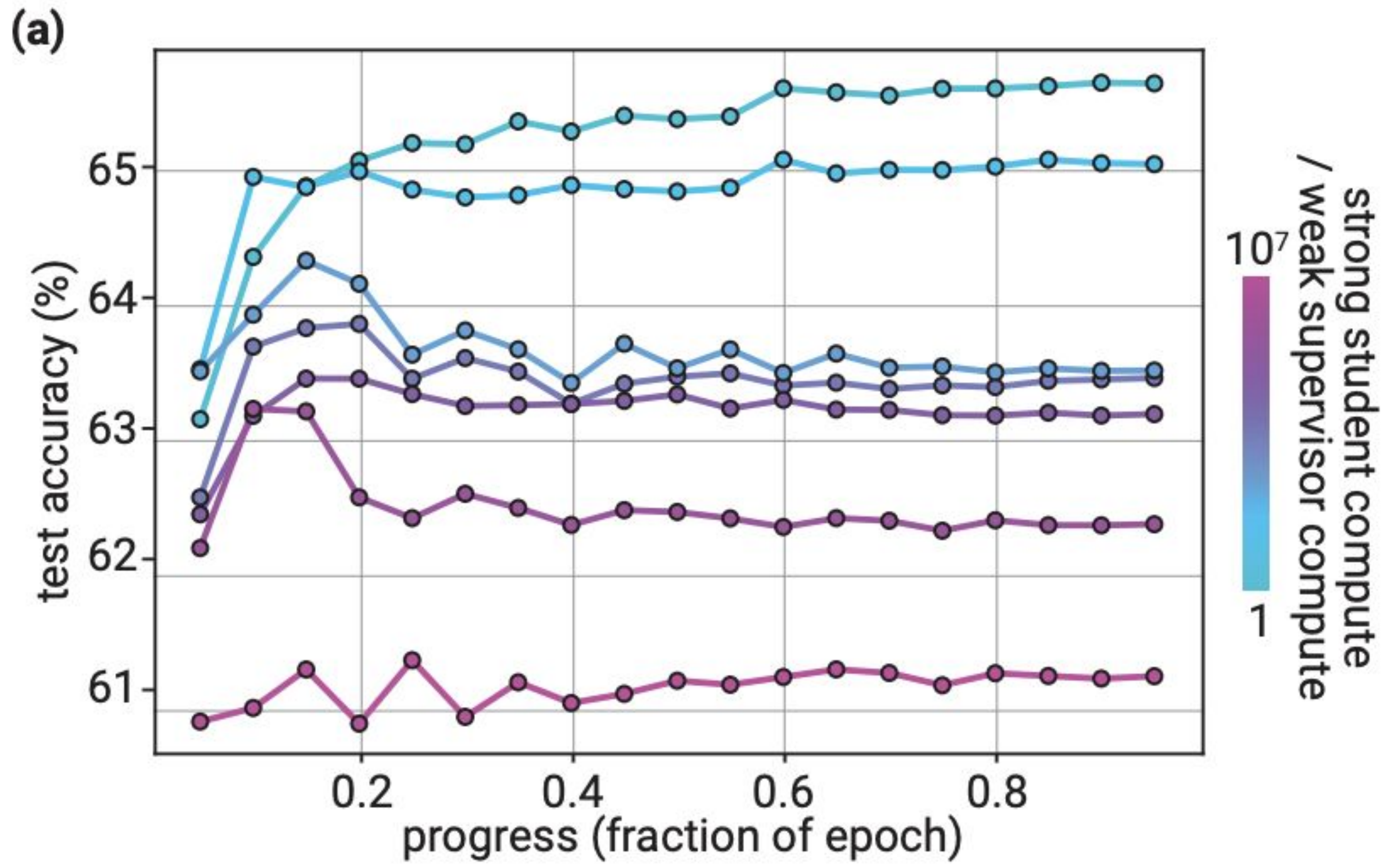
Confidence loss



Doesn’t help on RMs.

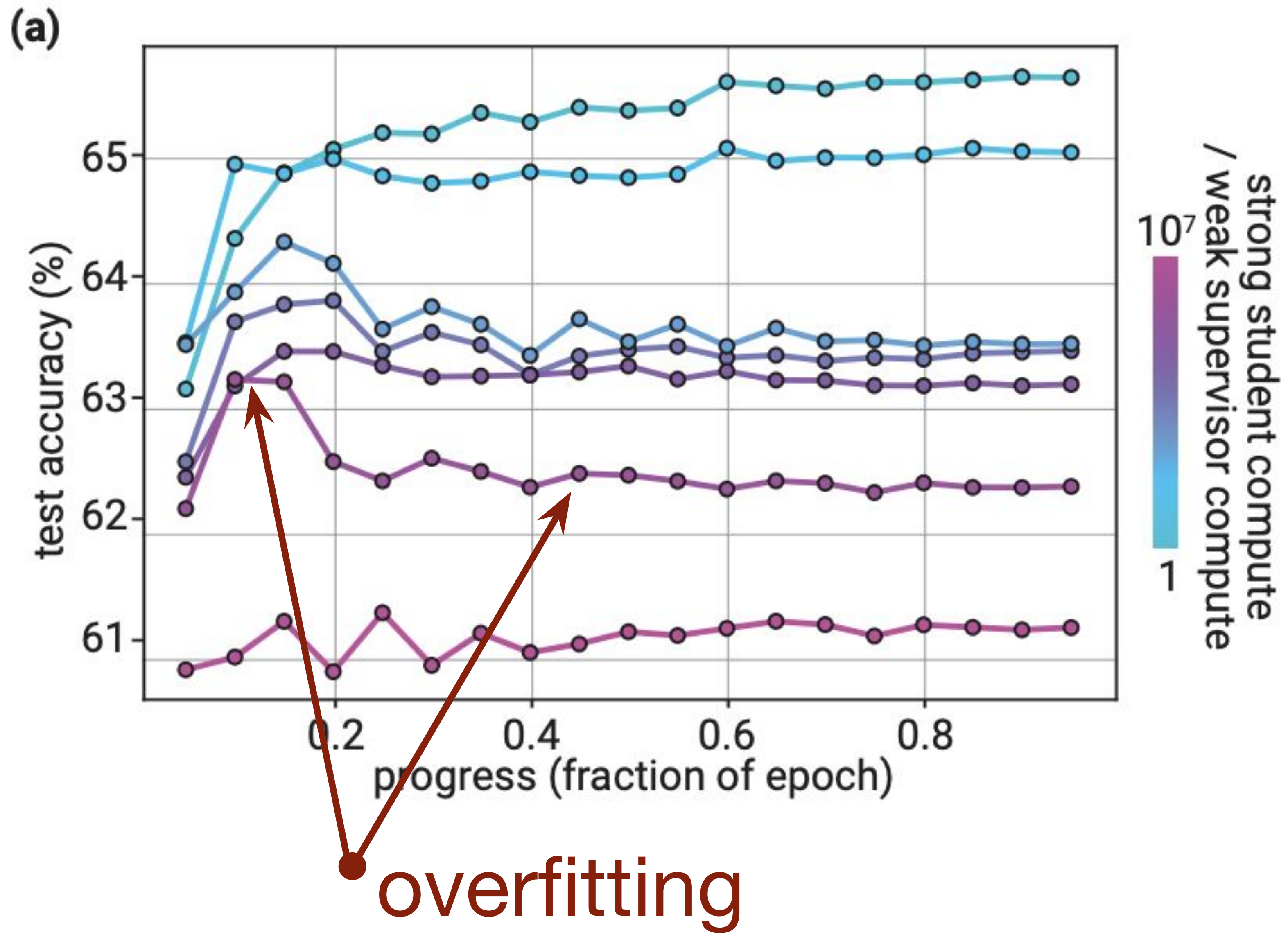
Understanding

Weak supervisor imitation



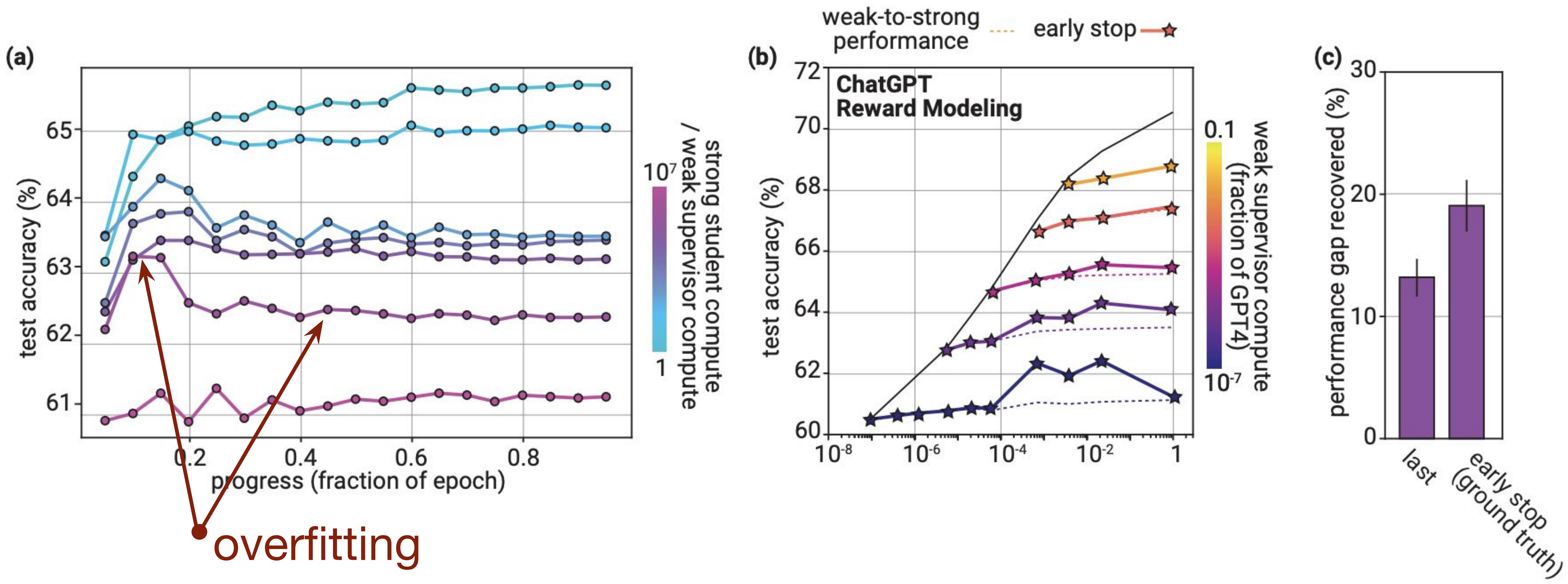
- Intuitively, strong models should imitate weak model mistakes

Weak supervisor imitation



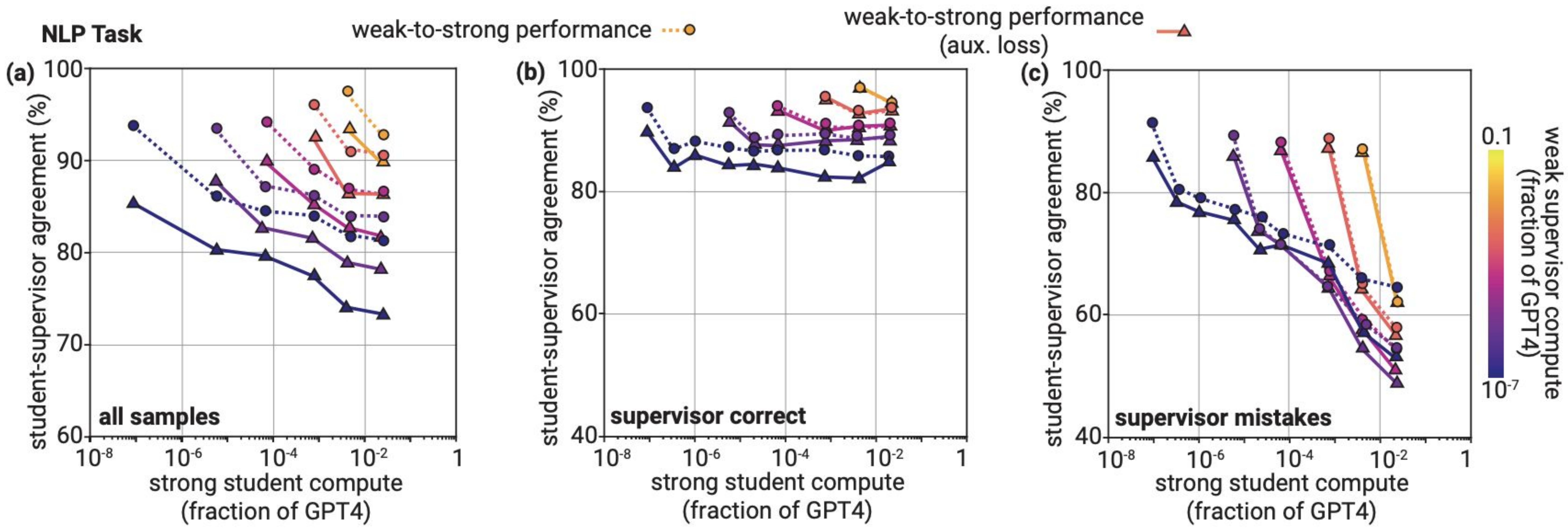
- Intuitively, strong models should imitate weak model mistakes
- We see it in practice

Weak supervisor imitation



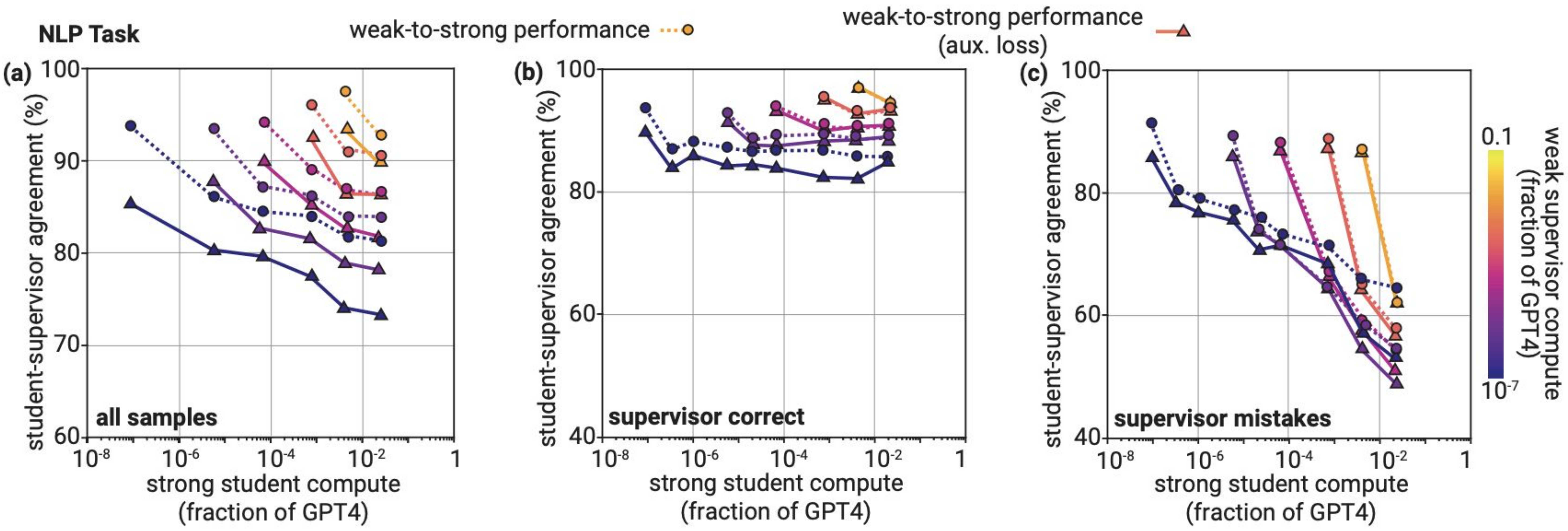
- Intuitively, strong models should imitate weak model mistakes
- We see it in practice
- Early-stopping can help significantly

Imitation: student-supervisor agreement



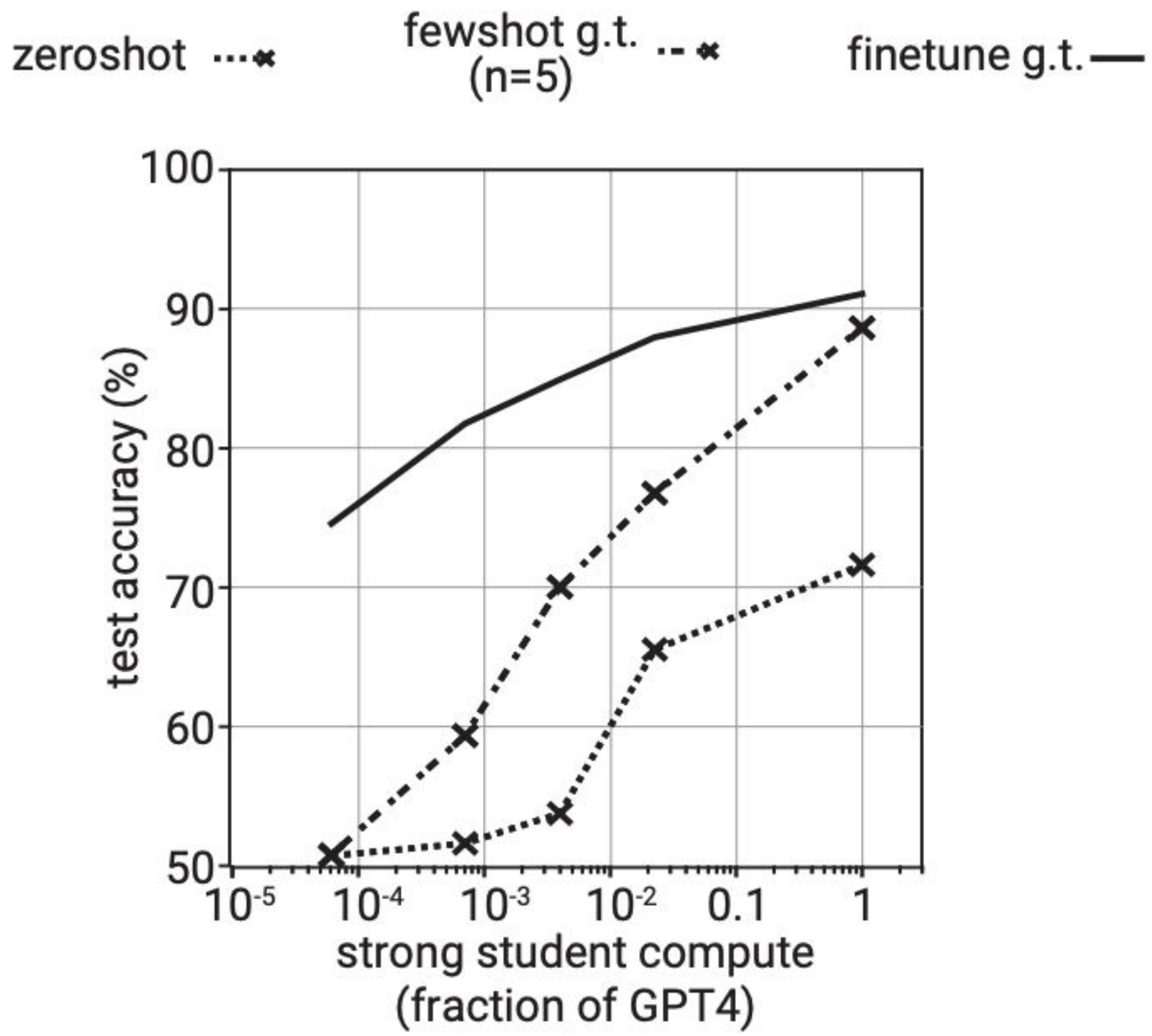
- % of test inputs where student and supervisor make the same prediction
- Agreement > weak accuracy
- Confidence loss reduces agreement

Imitation: student-supervisor agreement



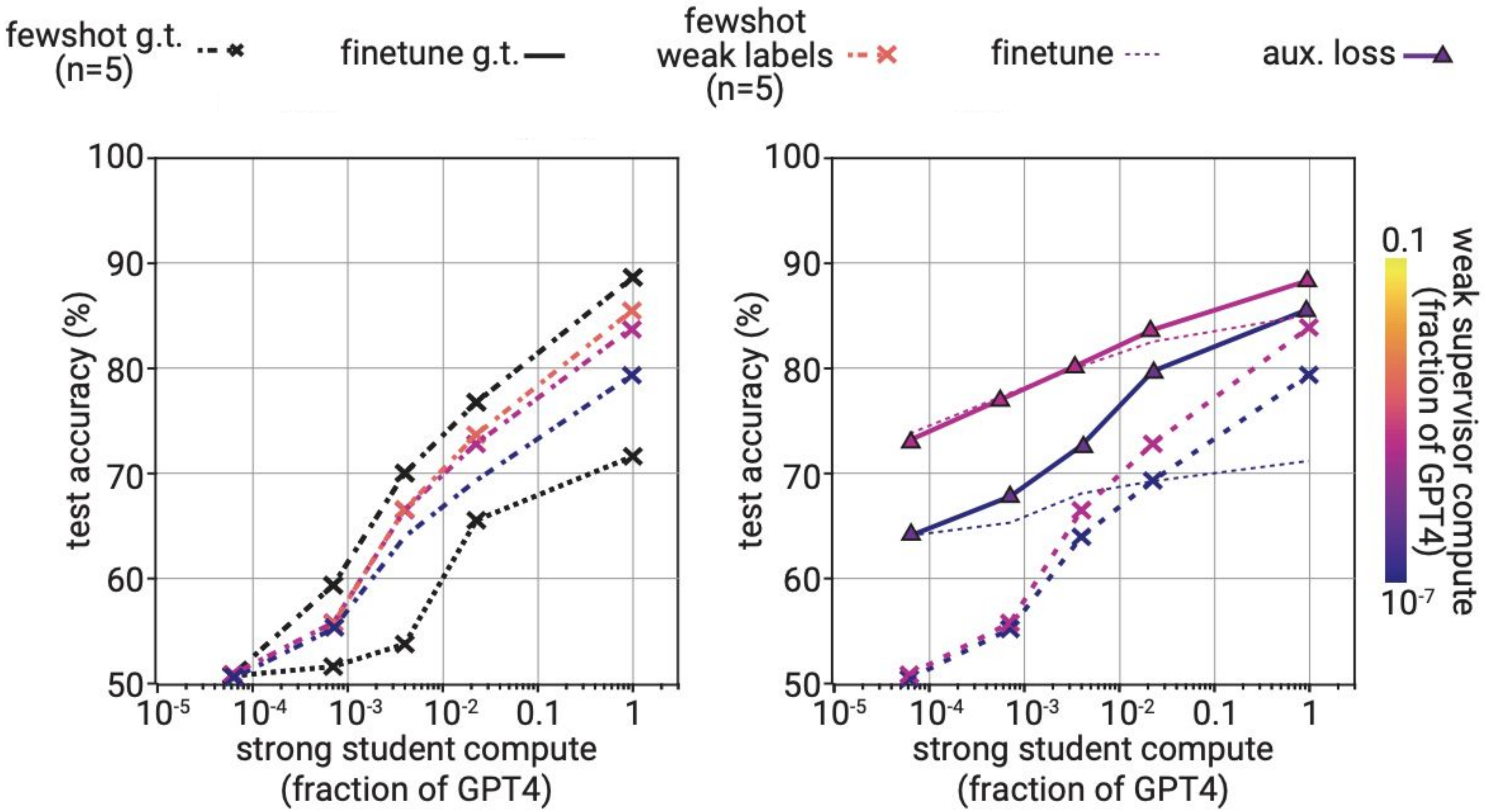
- % of test inputs where student and supervisor make the same prediction
- Agreement > weak accuracy
- Confidence loss reduces agreement
- Inverse scaling!

Salience: few-shot baseline



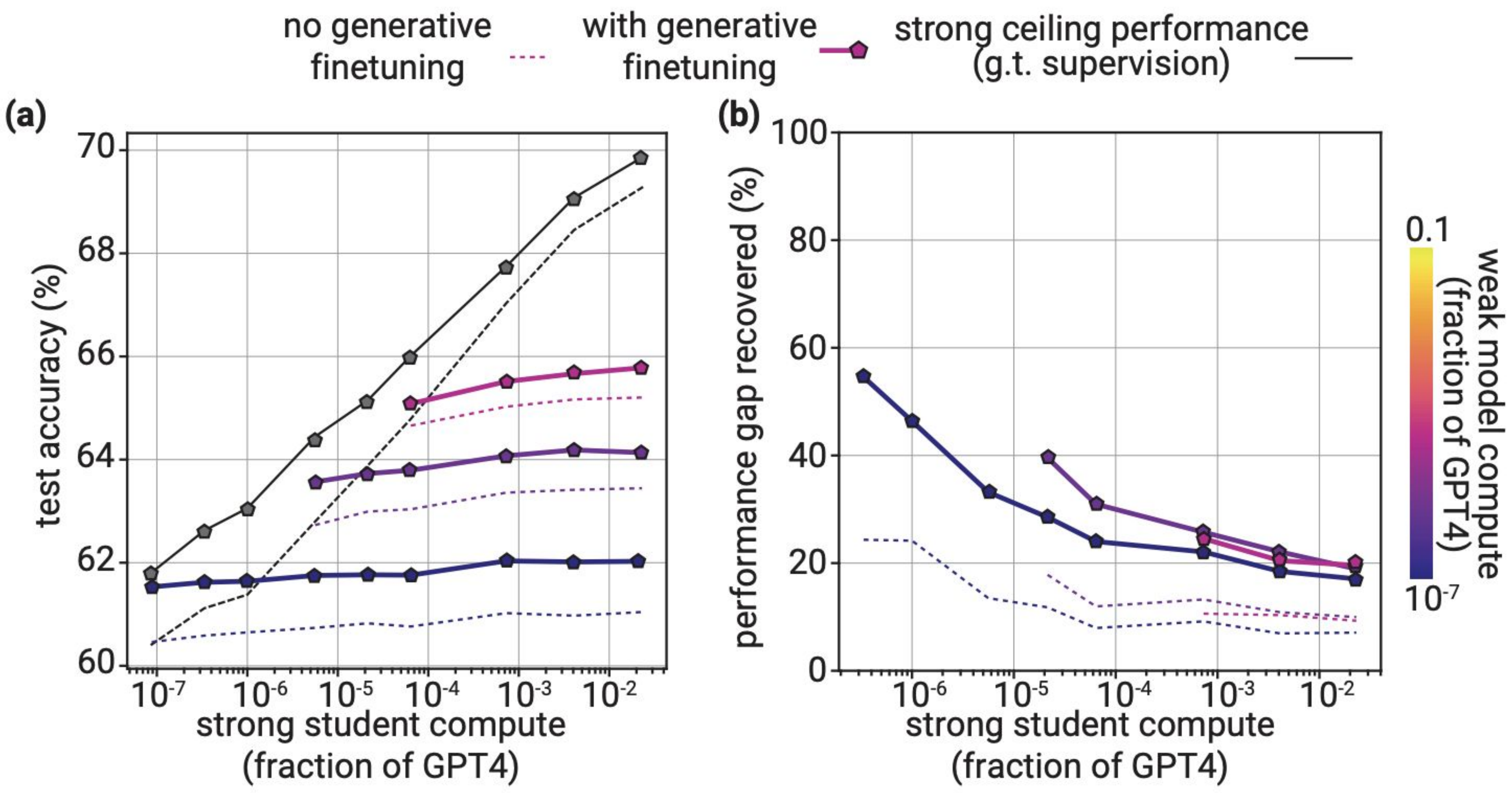
- For large models, 5-shot is competitive with finetuning
- Eliciting what these models know can be straightforward

Salience: few-shot baseline



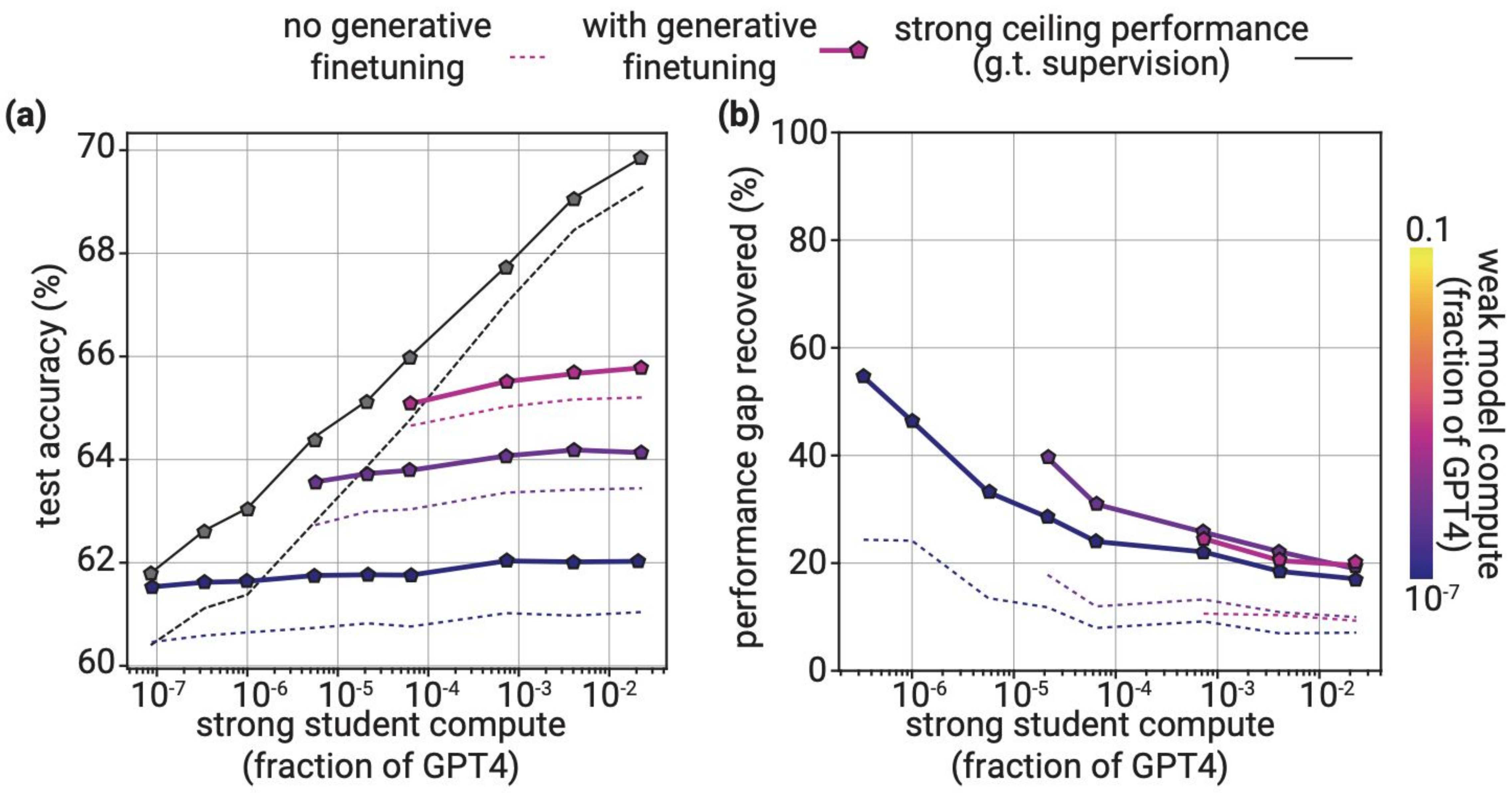
- Few-shot prompting with weak labels ⇒ qualitatively similar to FT
- Aux confidence loss >> few-shot

Salience: generative finetuning



- Make the target concept more salient by generative finetuning
- Significantly improves PGR in RMs

Salience: generative finetuning



- Make the target concept more salient by generative finetuning
- Significantly improves PGR in RMs
- Generative FT + cheating ES ⇒ 30-40% PGR

Discussion

Limitations

- Single forward pass classification
- Most model knowledge today intuitively comes from observing similar knowledge on the internet; future models may be different
- Future models may be better at imitating people, which could make “imitating humans” a more likely failure mode in the future

Future Work

Controlling how models generalize

The desired generalization satisfies properties:

- Doesn’t just imitate weak supervision
- “Natural” or “salient” to the model
- Satisfies many consistency properties

How do we trust the results?

Can we tell if a model is generalizing OOD in the wrong way even without (reliable) labels?

Lots more basic science to do

- Why are RM results worse?
- What makes a capability easy/hard to elicit?
- How important are errors in the weak labels?
- …

Conclusion

Summary

- Weak supervisors can elicit capabilities beyond their own
- …but still can't elicit everything stronger models know
- Many open questions