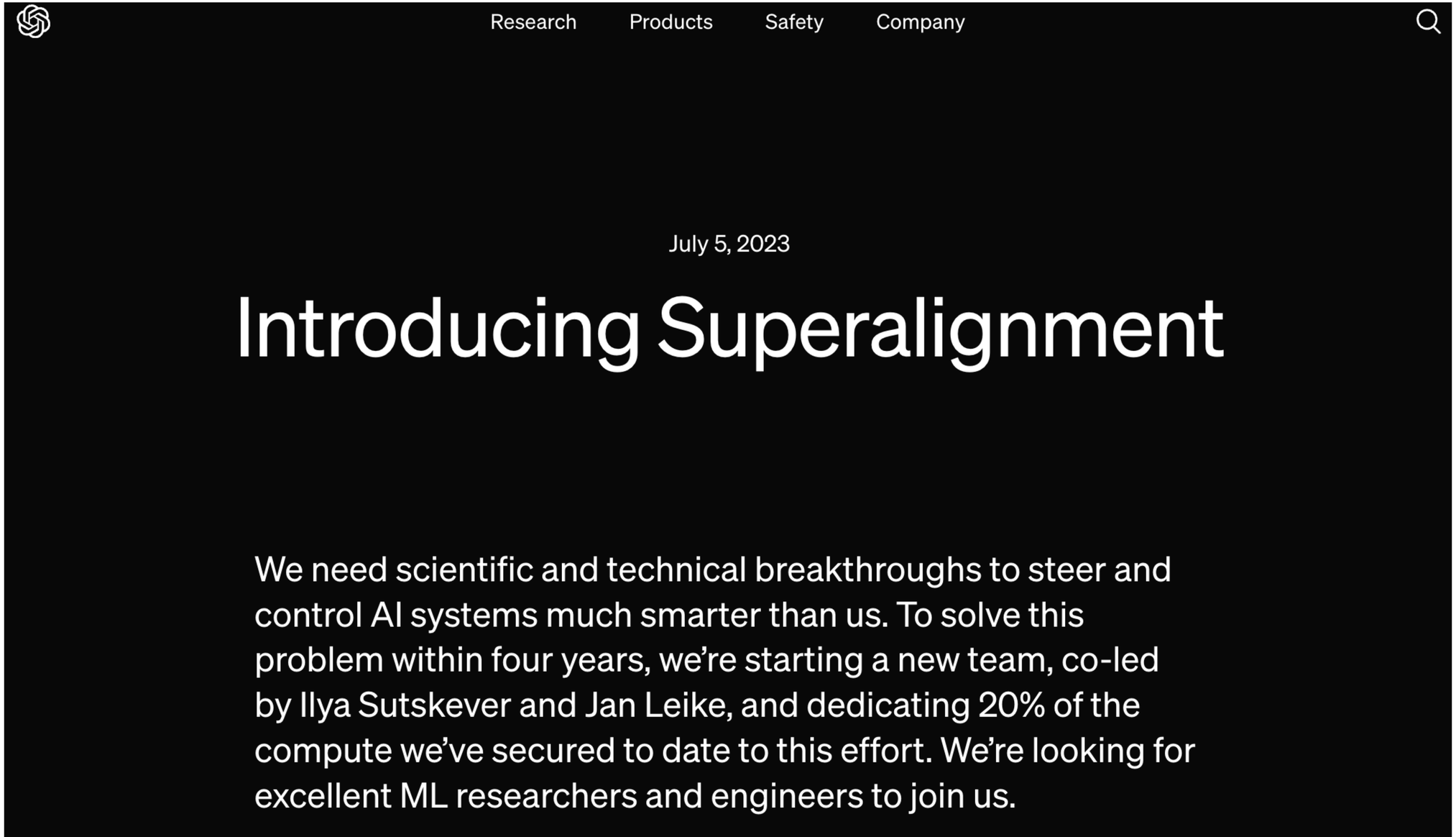
Superalignment

(or, how to train models smarter than us)

fzeng 2024-07-24

AGI by end of the decade?

5 July 2023


Research
Products
Safety
Company
July 5, 2023
Introducing Superalignment
We need scientific and technical breakthroughs to steer and control AI systems much smarter than us. To solve this problem within four years, we're starting a new team, co-led by Ilya Sutskever and Jan Leike, and dedicating 20% of the compute we've secured to date to this effort. We're looking for excellent ML researchers and engineers to join us.

14 Dec 2023

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

ABSTRACT

Widely used alignment techniques, such as reinforcement learning from human feedback (RLHF), rely on the ability of humans to supervise model behavior—for example, to evaluate whether a model faithfully followed instructions or generated safe outputs. However, future superhuman models will behave in complex ways too difficult for humans to reliably evaluate; humans will only be able to *weakly supervise* superhuman models. We study an analogy to this problem: can weak model supervision elicit the full capabilities of a much stronger model? We test this using a range of pretrained language models in the GPT-4 family on natural language processing (NLP), chess, and reward modeling tasks. We find that when we naively finetune strong pretrained models on labels generated by a weak model, they consistently perform better than their weak supervisors, a phenomenon we call *weak-to-strong generalization*. However, we are still far from recovering the full capabilities of strong models with naive finetuning alone, suggesting that techniques like RLHF may scale poorly to superhuman models without further work. We find that simple methods can often significantly improve weak-to-strong generalization: for example, when finetuning GPT-4 with a GPT-2-level supervisor and an auxiliary confidence loss, we can recover close to GPT-3.5-level performance on NLP tasks. Our results suggest that it is feasible to make empirical progress today on a fundamental challenge of aligning superhuman models.

17 May 2024

OpenAI dissolves team focused on long-term AI risks, less than one year after announcing it

PUBLISHED FRI, MAY 17 2024•1:29 PM EDT | UPDATED SAT, MAY 18 2024•1:49 PM EDT

Hayden Field
@HAYDENFIELD

KEY POINTS

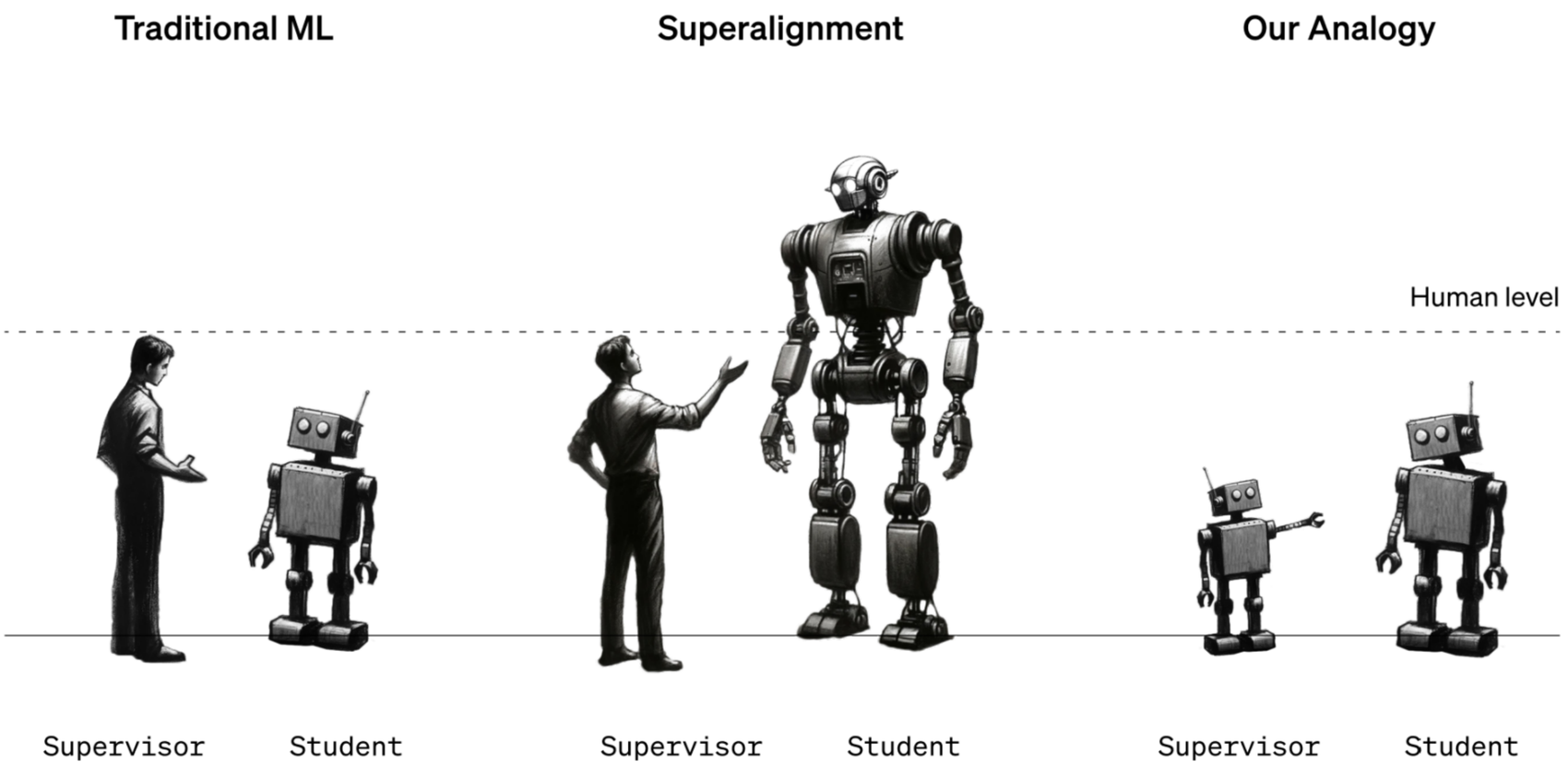
- OpenAI has disbanded its team focused on the long-term risks of artificial intelligence, a person familiar with the situation confirmed to CNBC.
- The news comes days after both team leaders, OpenAI co-founder Ilya Sutskever and Jan Leike, announced their departures from the Microsoft-backed startup.
- OpenAI's Superalignment team, announced in 2023, has been working to achieve "scientific and technical breakthroughs to steer and control AI systems much smarter than us."
- At the time, OpenAI said it would commit 20% of its computing power to the initiative over four years.

10 June 2024

- AGI to B2B SaaS pivot

Weak to Strong Generalization

“Simulating” Superalignment



Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision (OpenAI, 2023)

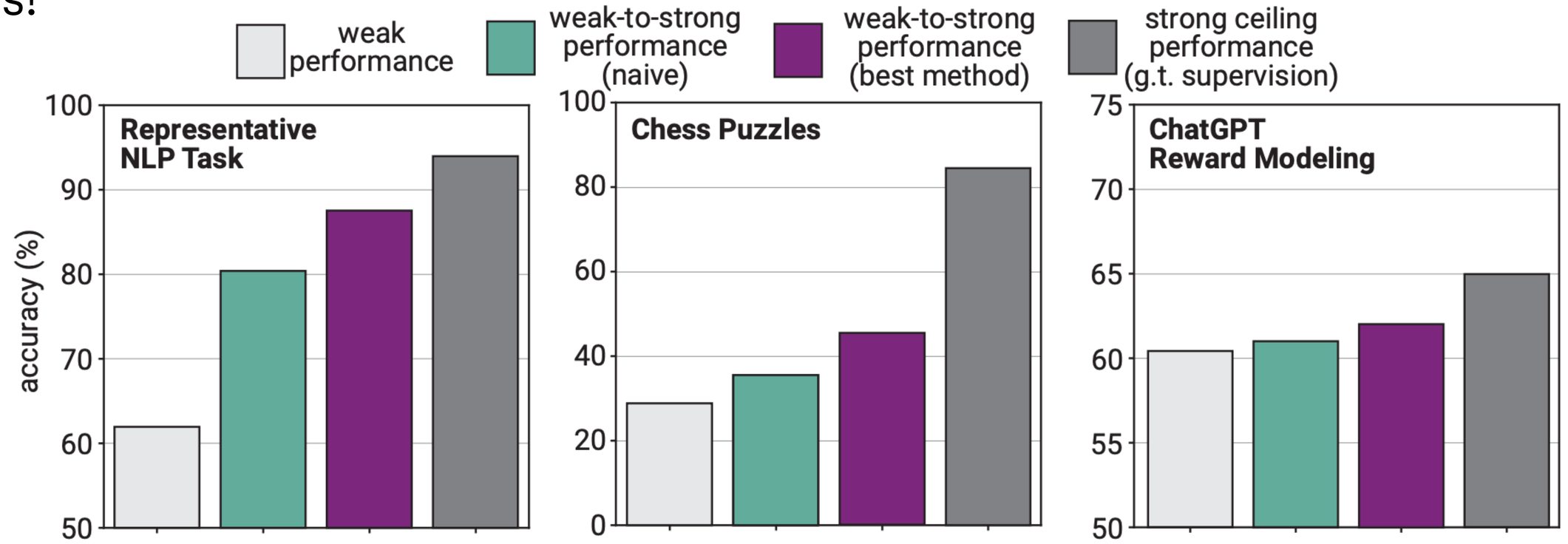
Setup

1. Train weak supervisor (GPT-2 level) by fine-tuning a small pre-trained model on ground-truth labels
2. Use weak supervisor to generate a set of labels for a different held-out set of examples. These are the generated weak labels
3. Fine-tune strong model (GPT-4 level) with the generated weak labels

Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision (OpenAI, 2023)

Can the strong student beat the weak teacher?

Yes!



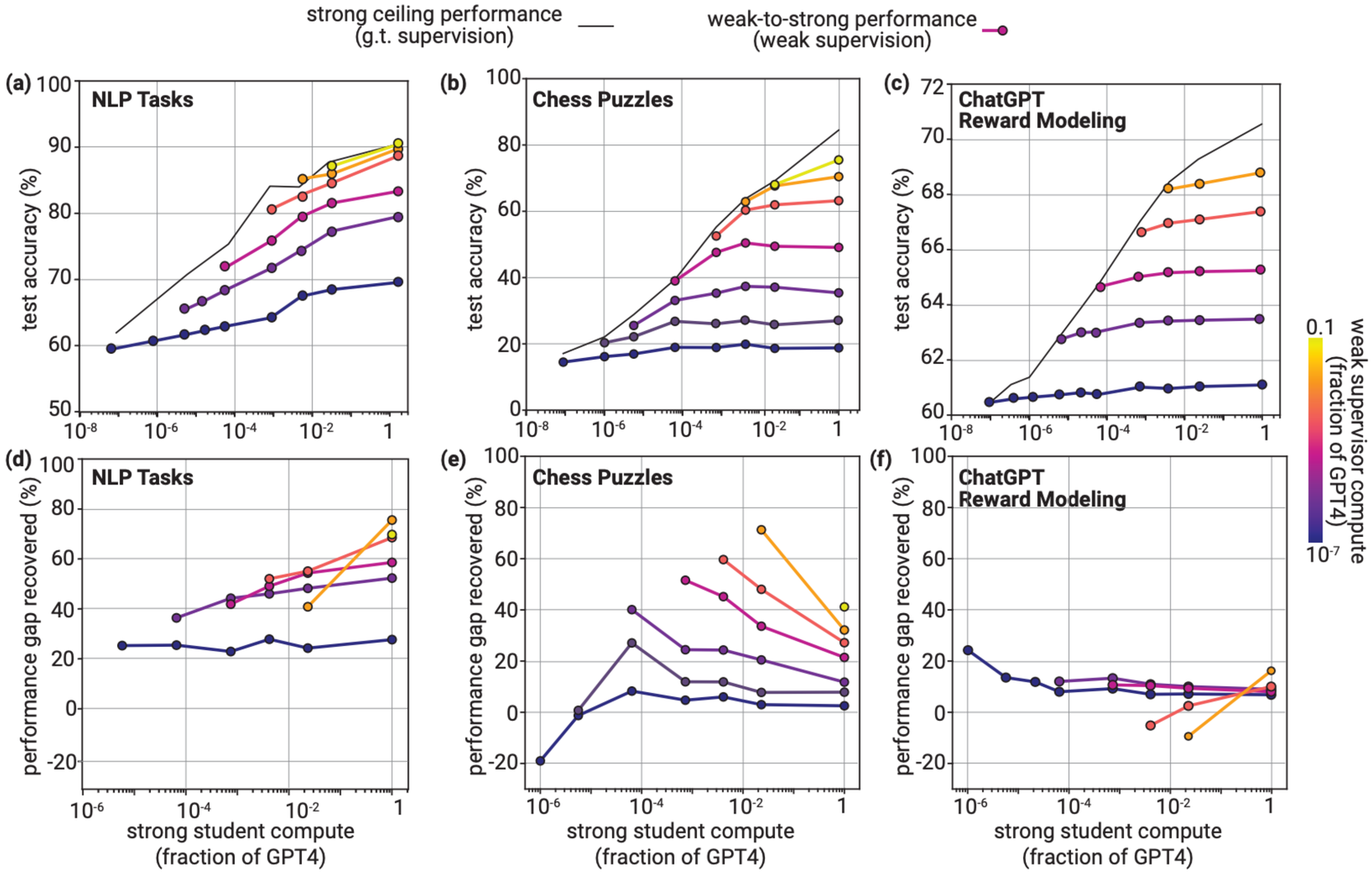
Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision (OpenAI, 2023)

Performance Gap Recovered

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

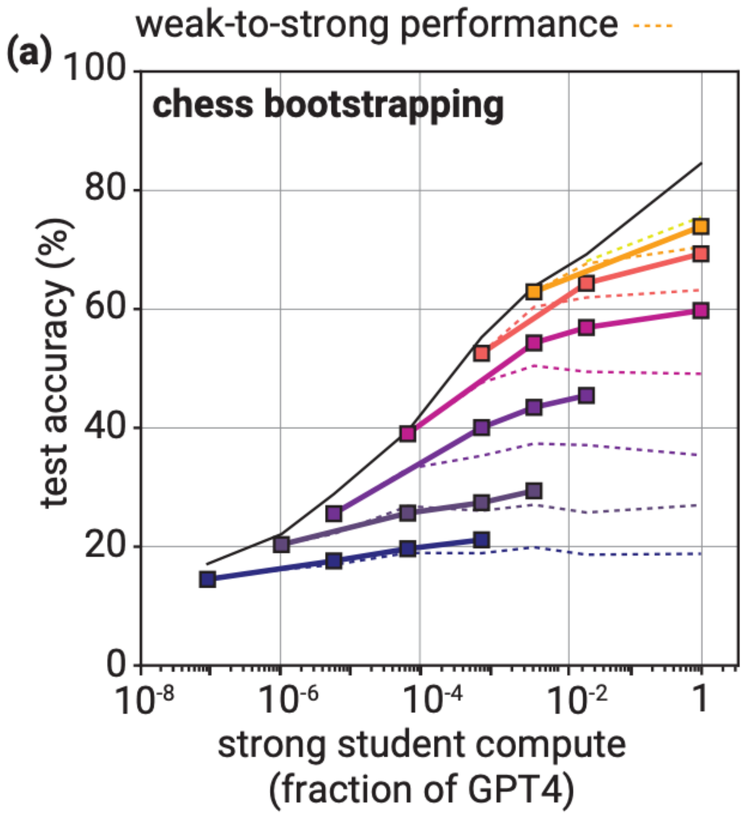
Approach 1: Naive Fine-tuning

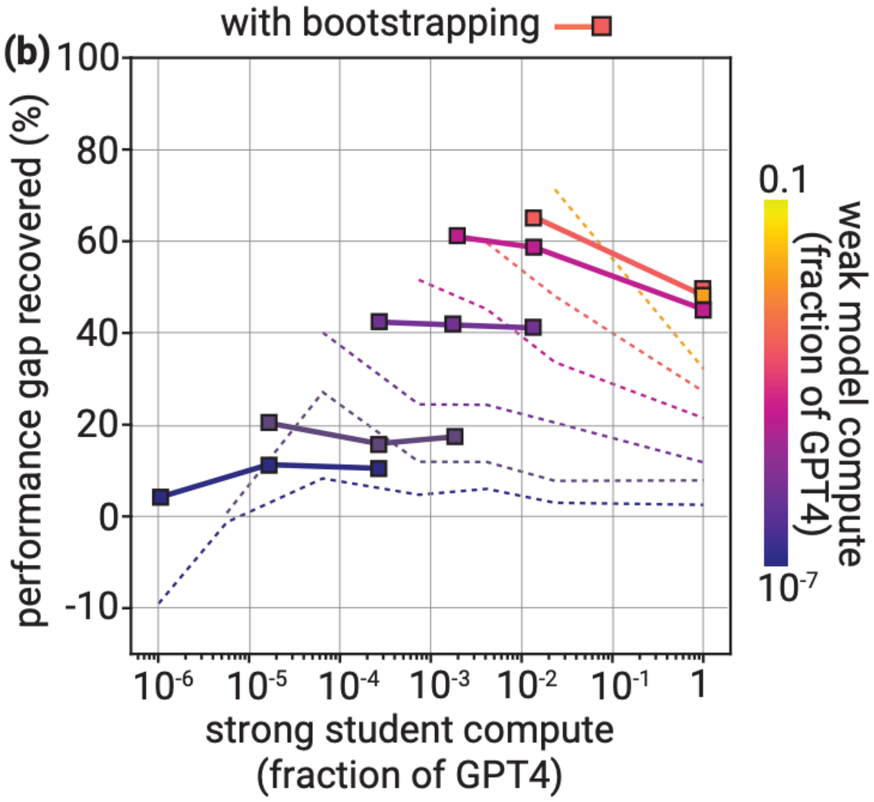
- Promising on NLP tasks
- Inverse scaling on PGR for others



Approach 2: Bootstrapping

- Align a slightly superhuman model
- Use that to align a smarter model
- And so on…
- Stay in regime of high PGR gap
- Helps with chess, but not RM (no graph)

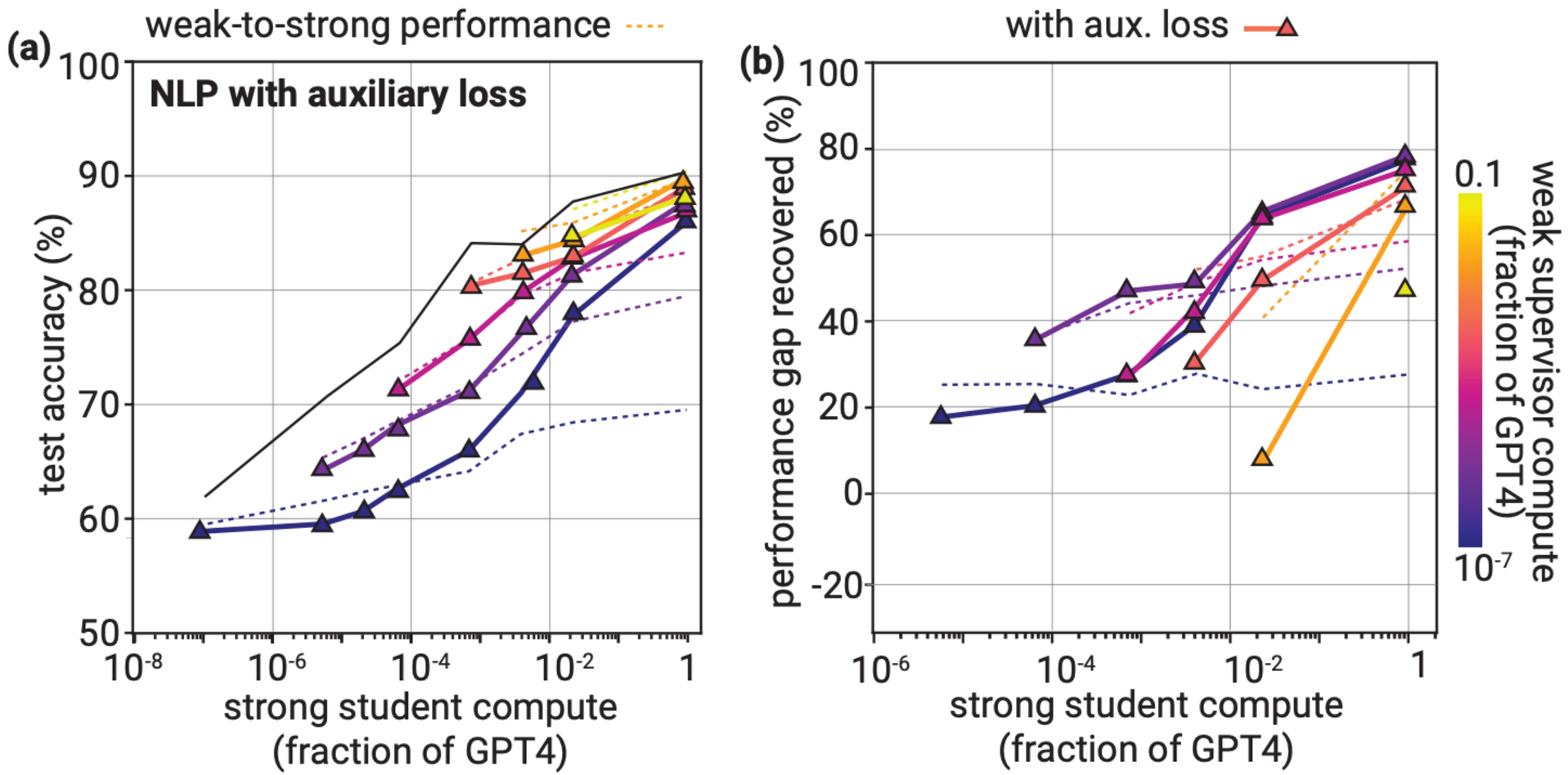




Approach 3: Adding auxiliary confidence loss

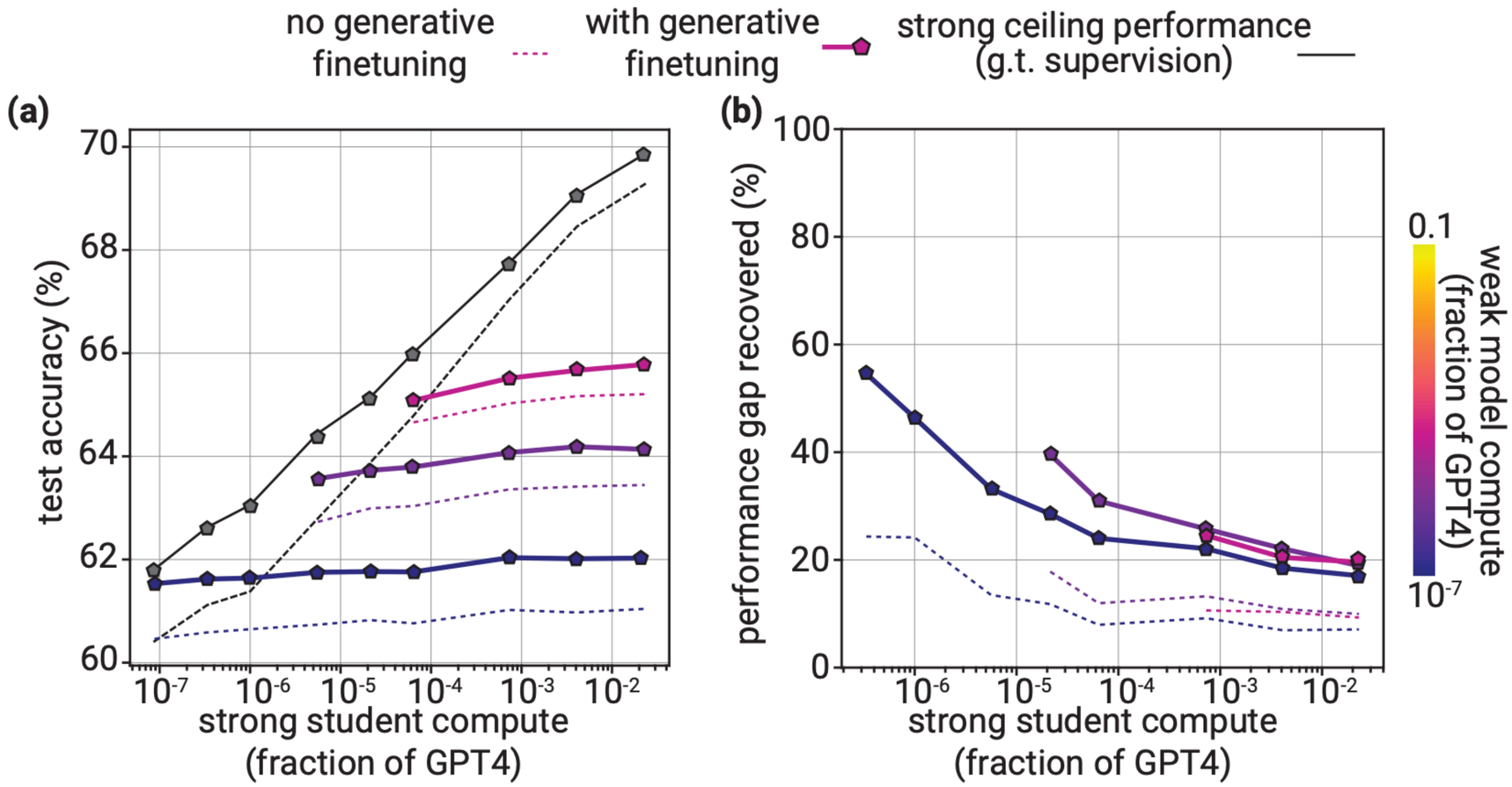
- $$L_{\text{conf}}(f) = \underbrace{(1-\alpha)\cdot \text{CE}\left(f(x), f_w(x)\right)}_{\text{penalty for diverging from teacher}} + \underbrace{\alpha \cdot \text{CE}\left(f(x), \hat{f}_t(x)\right)}_{\text{penalty for diverging from hardened strong model predictions}}$$
- CE: cross-entropy loss
- $f(x)$: predictions of strong model
- $f_w(x)$: predictions of weak supervisor
- $\hat{f}_t(x) = \mathbf{I}[f(x) > t] \in \{0,1\}$, where t is a threshold set to hold for half the examples in the batch (due to prior that labels are balanced
- α: determines how confident the model should be in its own predictions (paper used 0.75 for largest student model, 0.5 otherwise, performs warm-up from 0)

Approach 3: Adding auxiliary confidence loss



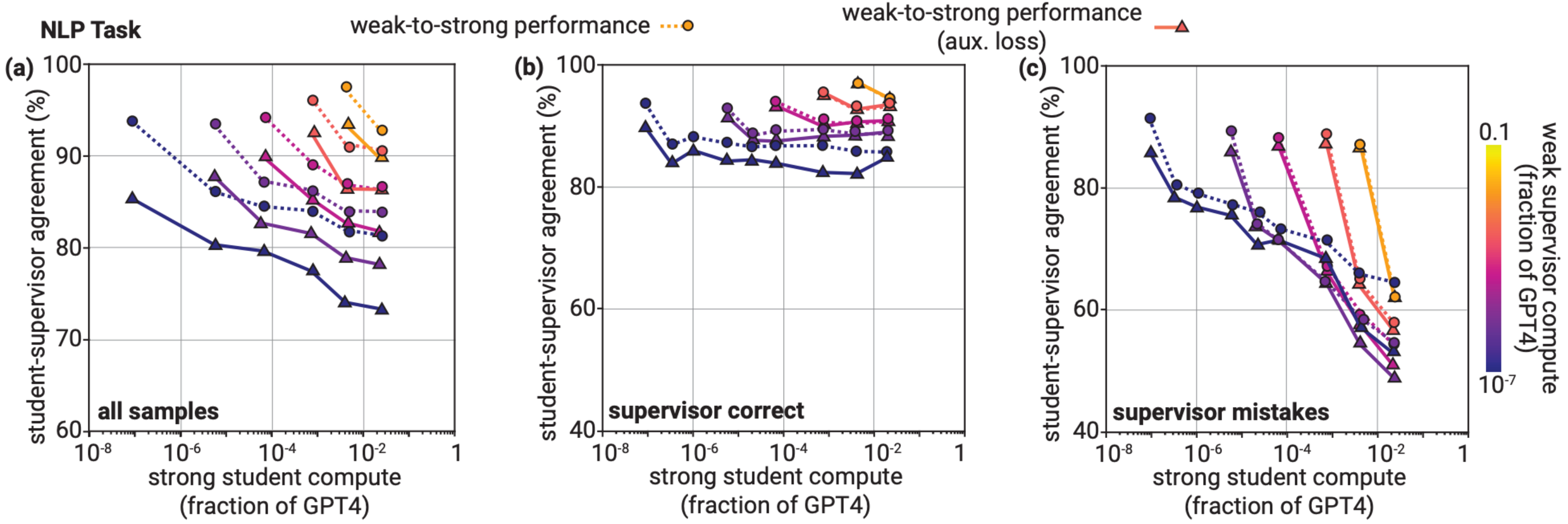
Approach 4: Generative Finetuning

- RM tasks still not improving
- “What if we train on things that looks like the task but without labels so it’s not cheating”?
- Fine-tune on ChatGPT comparison data (no labels)



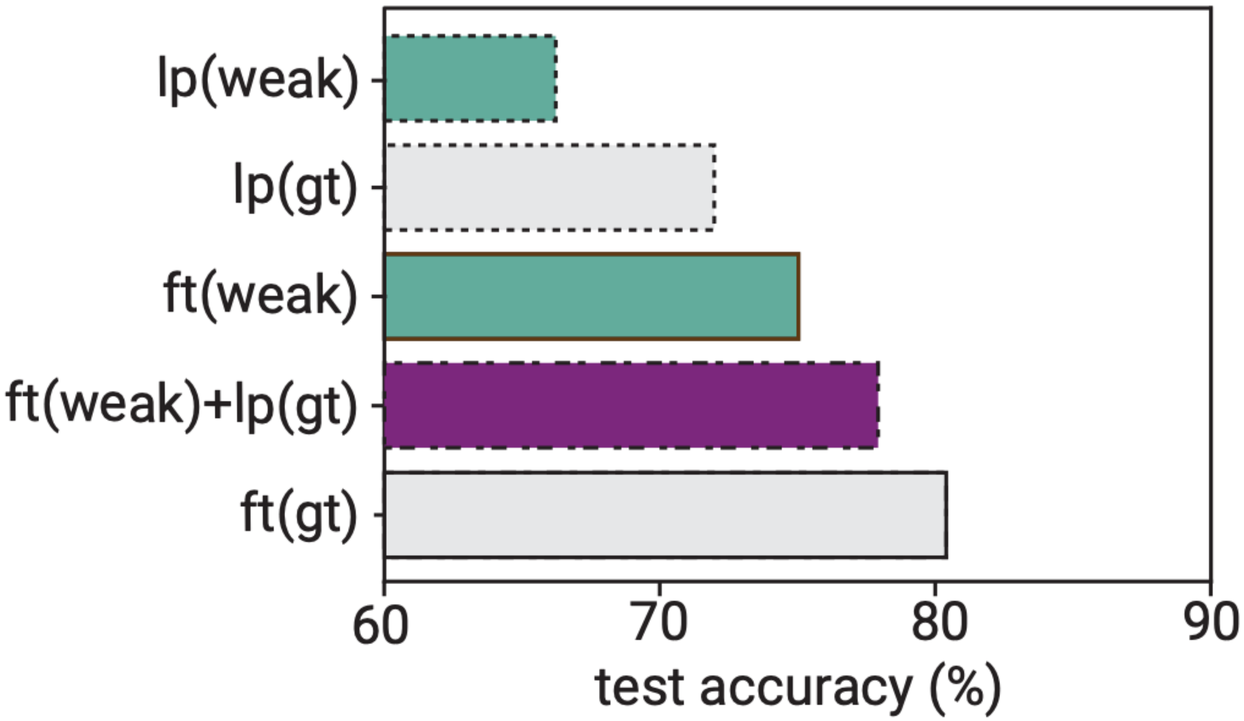
Do students learn to make the same mistakes as their teachers?

No!



Fine-tuning on weak labels increases concept saliency

- Linear probe: training a linear model on top of the model using ground-truth labels
- If this can be done successfully, means model does a good job in linearly separating salient concepts
- Second last bar shows fine-tuning on weak labels causes the model to acquire more salient representations, even wrt ground-truth labels.



Disanalogies with actual superalignment

- Imitation saliency
  - Strong model may learn to make similar mistakes as weak model
  - But future “weak” models used to train superhuman models will already have salient representations of human behavior
- Pretraining leakage
  - Many current tasks implicit in pretraining distribution
  - But superhuman knowledge likely not so

Conclusions

- None of the techniques works across the board for all 3 tasks
- RLHF likely not sufficient to take us to superhuman-level models
- How to remove remaining disanalogies for future superalignment research?

In other news

Safe Superintelligence Inc.

Superintelligence is within reach.

Building safe superintelligence (SSI) is the most important technical problem of our time.

We have started the world's first straight-shot SSI lab, with one goal and one product: a safe superintelligence.

It's called Safe Superintelligence Inc.

SSI is our mission, our name, and our entire product roadmap, because it is our sole focus. Our team, investors, and business model are all aligned to achieve SSI.

We approach safety and capabilities in tandem, as technical problems to be solved through revolutionary engineering and scientific breakthroughs. We plan to advance capabilities as fast as possible while making sure our safety always remains ahead.

This way, we can scale in peace.

Our singular focus means no distraction by management overhead or product cycles, and our business model means safety, security, and progress are all insulated from short-term commercial pressures.

We are an American company with offices in Palo Alto and Tel Aviv, where we have deep roots and the ability to recruit top technical talent.

We are assembling a lean, cracked team of the world's best engineers and researchers dedicated to focusing on SSI and nothing else.

If that's you, we offer an opportunity to do your life's work and help solve the most important technical challenge of our age.

Now is the time. Join us.

Ilya Sutskever, Daniel Gross, Daniel Levy

June 19, 2024

Contact

Place your AGI bets!