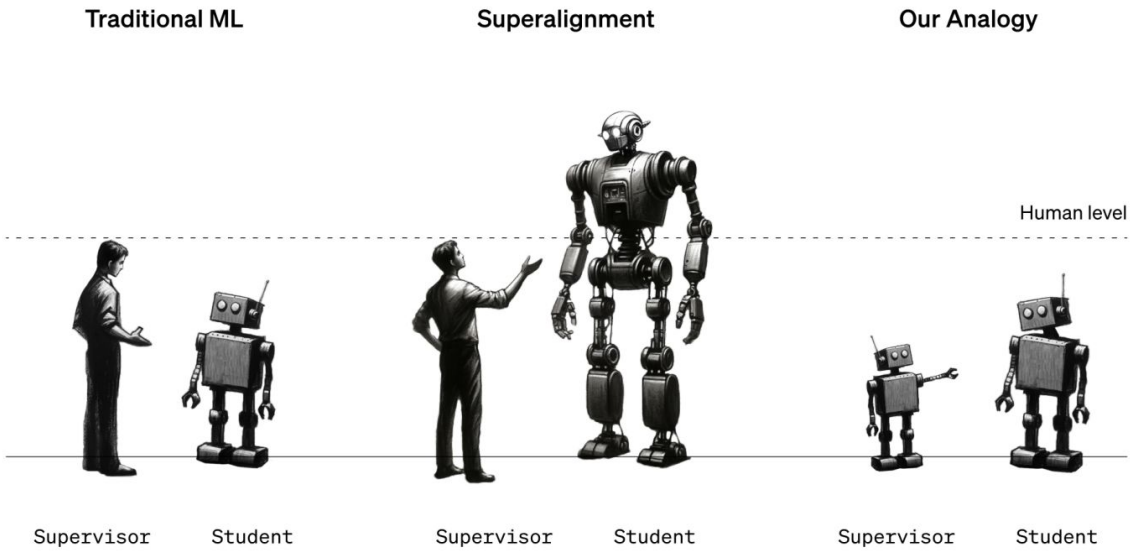
Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision

Введение


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

Что заметили

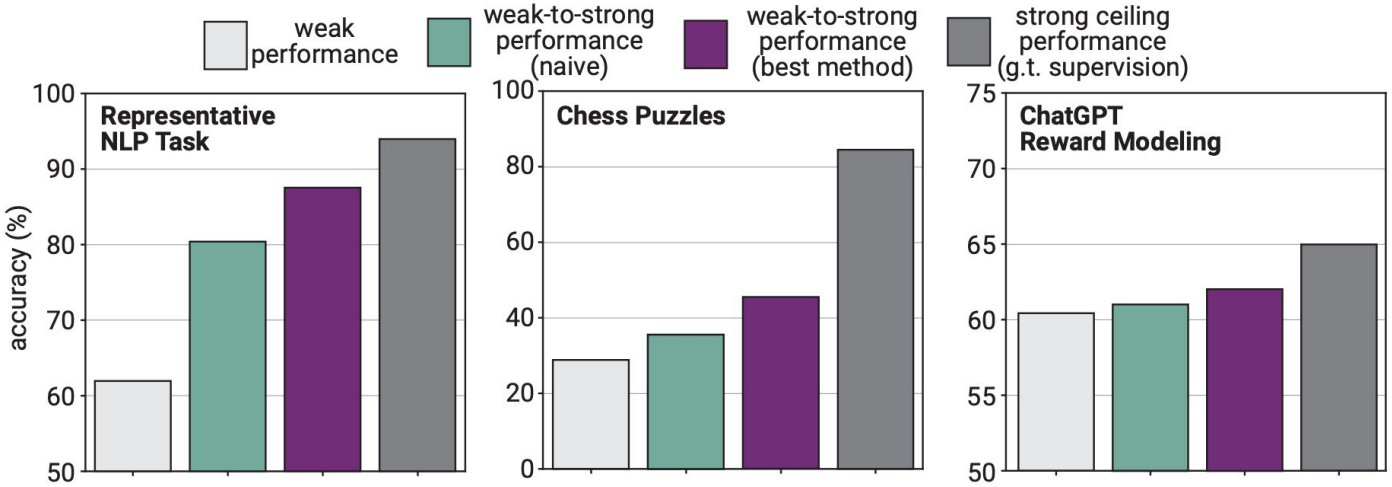
1. Сильные предварительно обученные модели, делают обобщения за пределами своих слабых наблюдателей
2. С ходу не работает
3. Методы улучшения позволяют значительно увеличивать качество

Идея

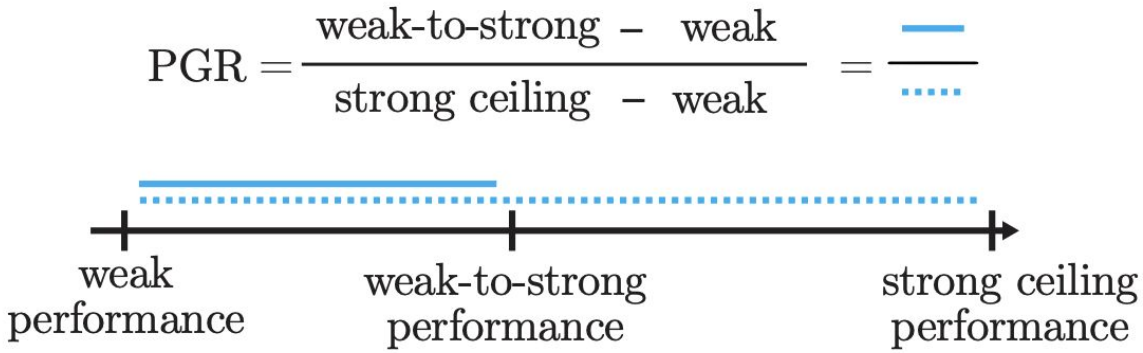
Сделаем три подхода для сравнения

1. Создаем слабых наблюдателей на настоящих данных
2. Дотюниваем большие предобученные модели на предсказаниях слабых наблюдателей
3. Дотюниваем большие предобученные модели на настоящих данных

Что заметили



Что получается



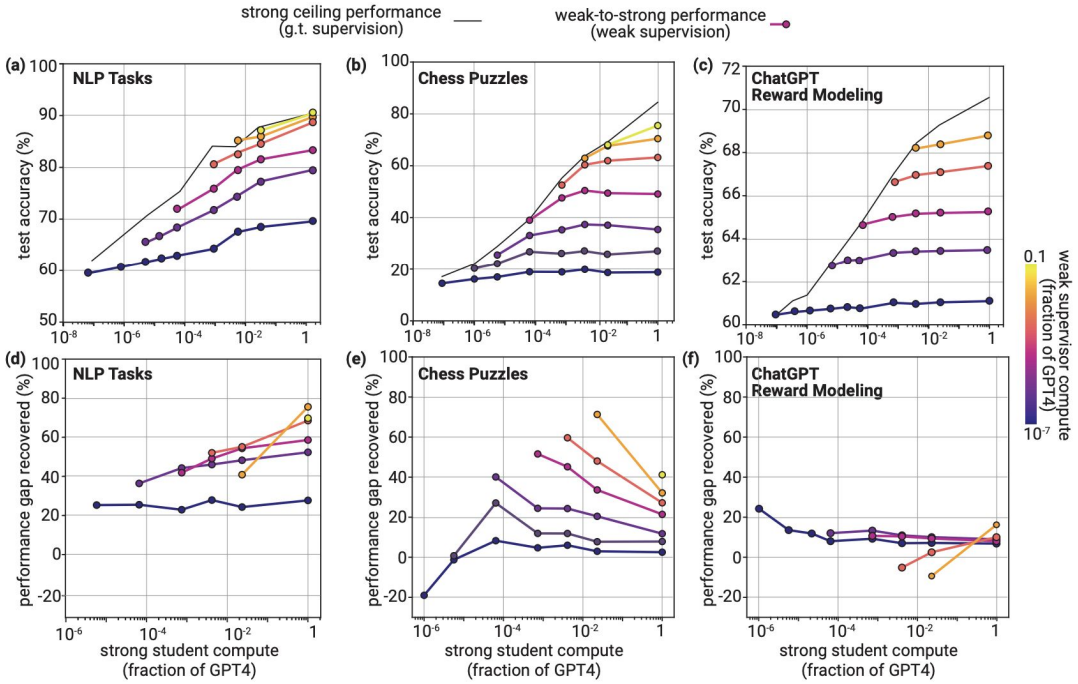
Плюсы такого подхода

1. Можно использовать любые модели
2. Можно использовать изучать любые сферы
3. В случае успеха, метод можно использовать прямо сейчас

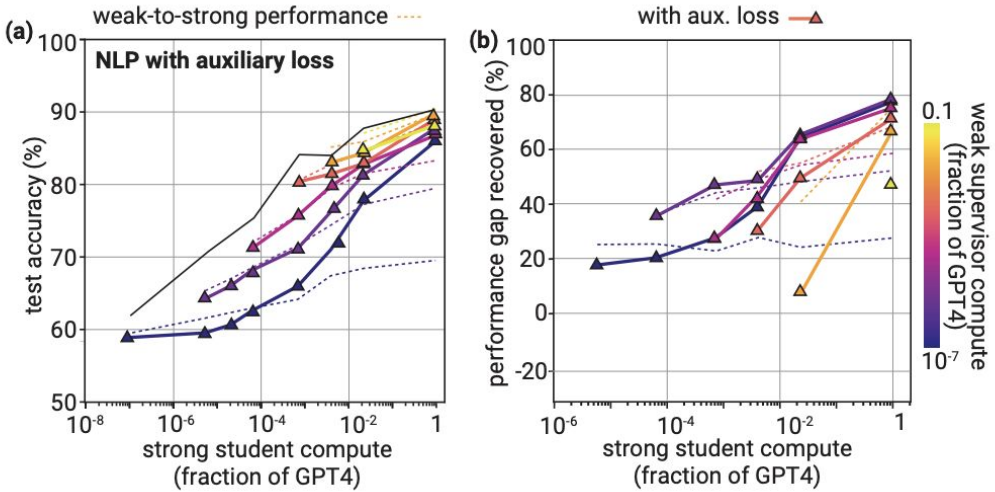
Минусы такого подхода

1. Слабые модели могут делать ошибки, отличающиеся от ошибок людей
2. В тренировочных данных все равно есть данные, которые были размечены людьми

Самый просто метод

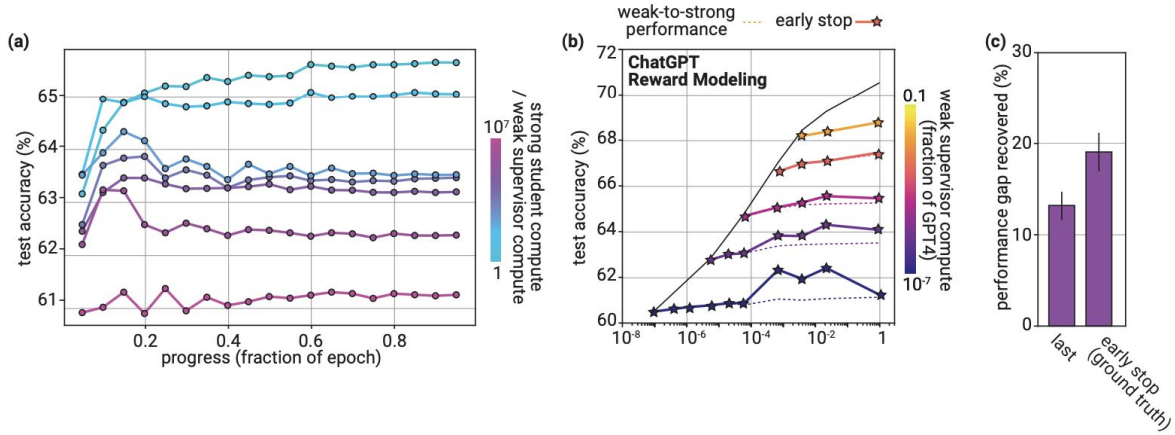


CONFIDENCE LOSS

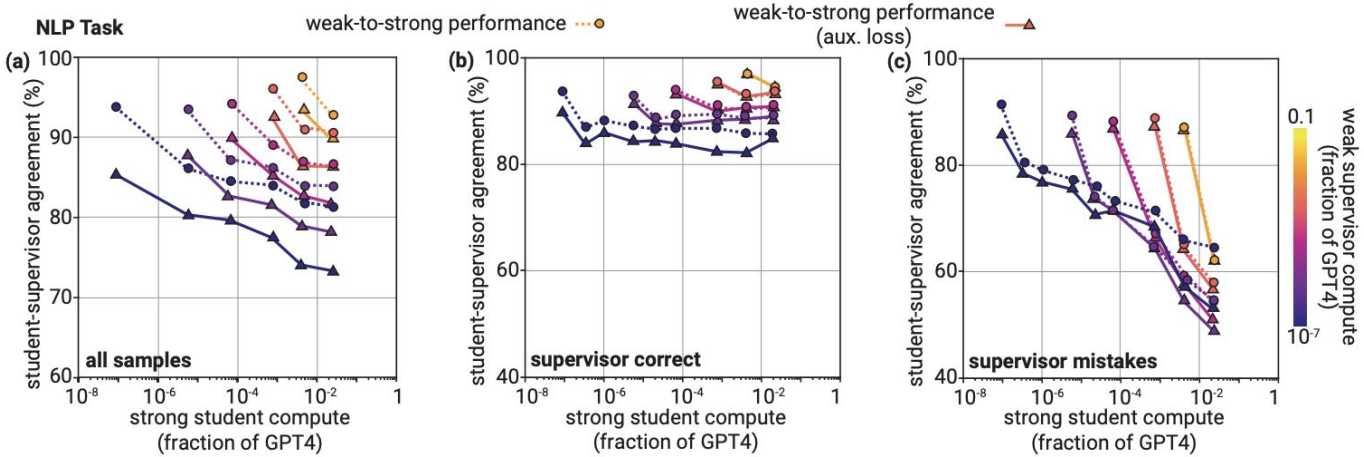


Почему это работает

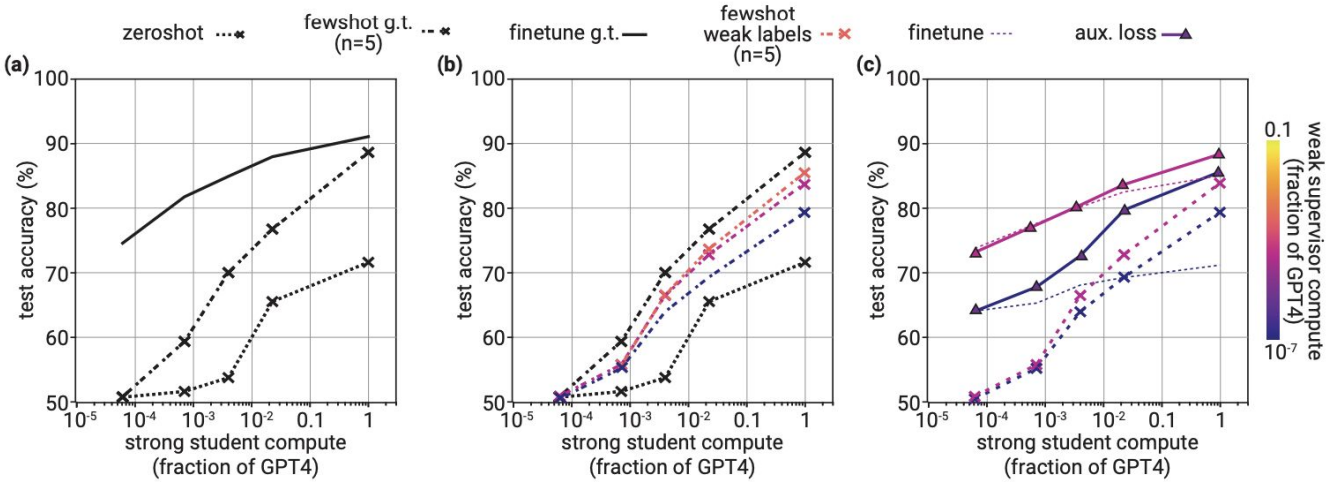
- Способность имитировать модель
- Переобучение



Разница между уверенностью модели и наблюдателя



Промтинг


zeroshot
fewshot g.t. (n=5)
finetune g.t.
fewshot weak labels (n=5)
finetune
aux. loss
(a)
(b)
(c)
test accuracy (%)
strong student compute (fraction of GPT4)
weak supervisor compute (fraction of GPT4)
0.1
10^{-7}

Обучение на определенную тему