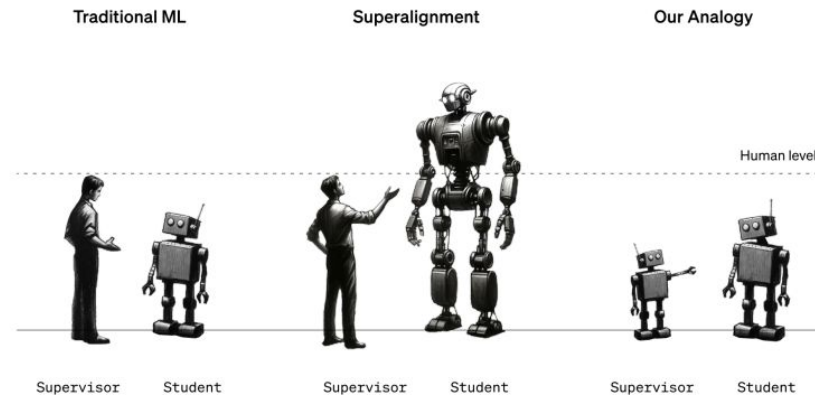


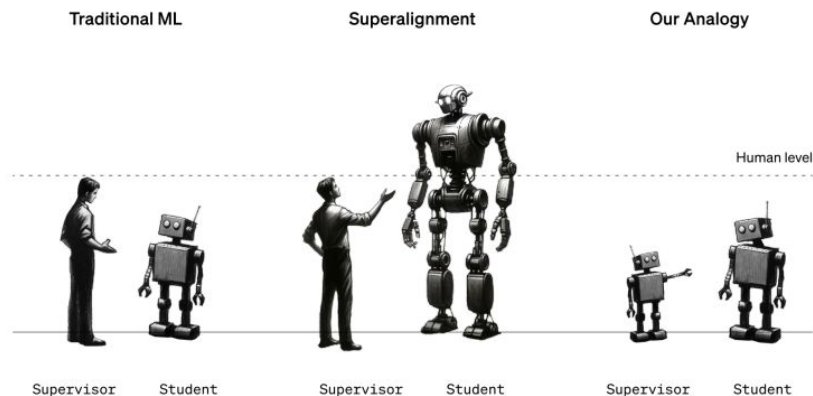
# **WEAK-TO-STRONG GENERALIZATION: ELICITING STRONG CAPABILITIES WITH WEAK SUPERVISION**

Подготовила Жумлякова Светлана, БПМИ203

# Что такое weak-to-strong?



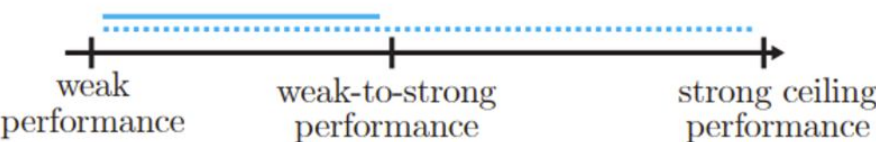
# Может ли weak-to-strong работать?



- Модель-ученик может начать копировать слабого учителя и его ошибки
- + Слабая модель поможет сильной модели “раскрыть свой потенциал”

# Пайплайн

- Модель-ученик GPT4 (GPT3.5), модель-учитель GPT2
- Обучается модель-учитель на ground truth
- Модель-учитель генерирует разметку
- Предобученная модель-ученик дообучается на разметке учителя (не ground truth)
- Обучается сильная модель на ground truth - как пример максимального уровня обучения
- Считается PGR

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


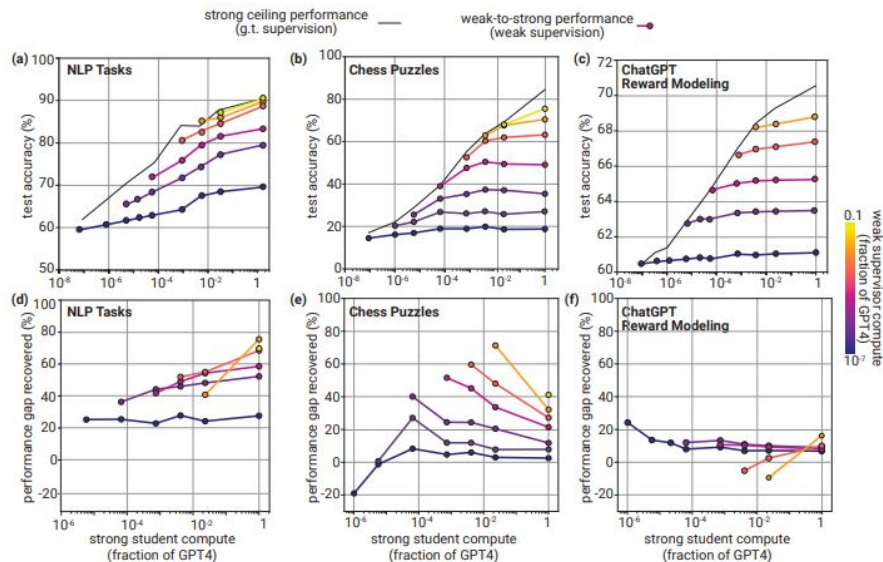
The diagram illustrates a performance scale on a horizontal axis with an arrow pointing right. Three points are marked on the axis: 'weak performance' on the left, 'weak-to-strong performance' in the middle, and 'strong ceiling performance' on the right. Above the axis, a solid blue horizontal line segment spans from the 'weak performance' point to the 'weak-to-strong performance' point. A dotted blue horizontal line segment spans from the 'weak performance' point to the 'strong ceiling performance' point. The solid line is positioned above the dotted line, visually representing the numerator of the PGR formula as a fraction of the total range.



# Задачи

- 22 NLP задачи (этика, настроение и т.п)
  - Данные -> бинарная классификация
- Шахматные задачи
  - Головоломки на поиск оптимальных ходов
  - По позициям фигур предсказать оптимальный ход
  - Генеративная задача
- Reward modeling
  - (dialog, ans1, ans2) -> 1 if ans1 better than ans2

# Результаты: наивный метод



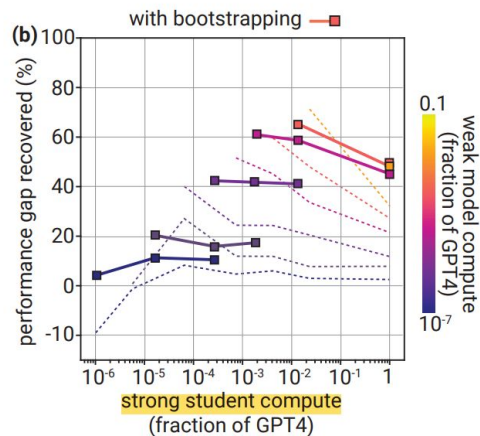
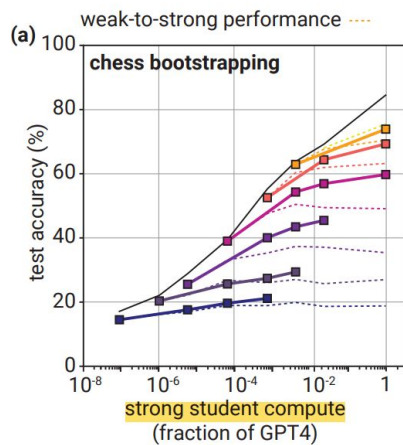
- PGR > 0 - ученик работает лучше учителя!
- NLP:
  - PGR > 20, растет с ростом размеров ученика и учителя
- Шахматы:
  - для слабого учителя PGR очень мал
  - падение качества при росте размера ученика
- RM:
  - PGR низок



## Улучшение 1: Бутстрэп

- $M_1 \dots M_n$  - модели; с ростом номера растет размер модели
- $M_1$  обучается на разметке учителя
- $M_1$  генерирует разметку
- $M_2$  обучается на разметке  $M_1$
- ...

# Улучшение 1: Бутстрэп



- Шахматы: рост PGR
- NLP и RM - без изменений





## Улучшение 2: auxiliary confidence loss

Зачем? Хотим, чтобы модель-ученик меньше запоминала ошибки модели-учителя

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

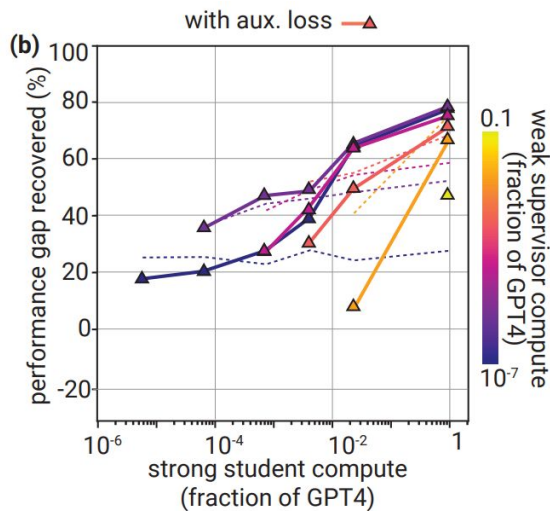
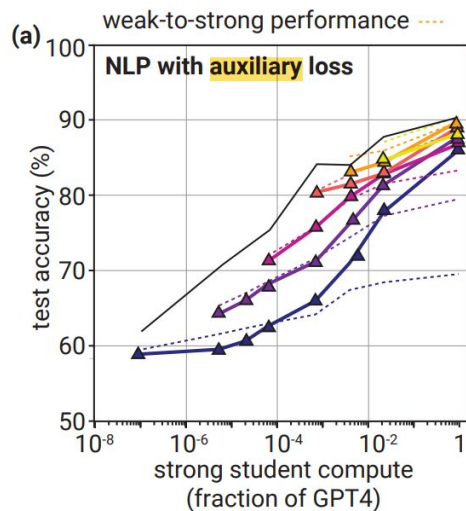
$f(x)$  in  $[0; 1]$  - the strong label predictive distribution

$f_w(x)$  in  $[0; 1]$  - the weak label predictive distribution

$t$  - threshold

$f_t(x) = \mathbb{I}[f(x) > t]$  in  $[0, 1]$  - the hardened strong model predictions using a threshold

## Улучшение 2: auxiliary confidence loss



- NLP
- Для маленьких размеров модели-ученика качество хуже
- Рост качества для моделей-учеников больших размеров

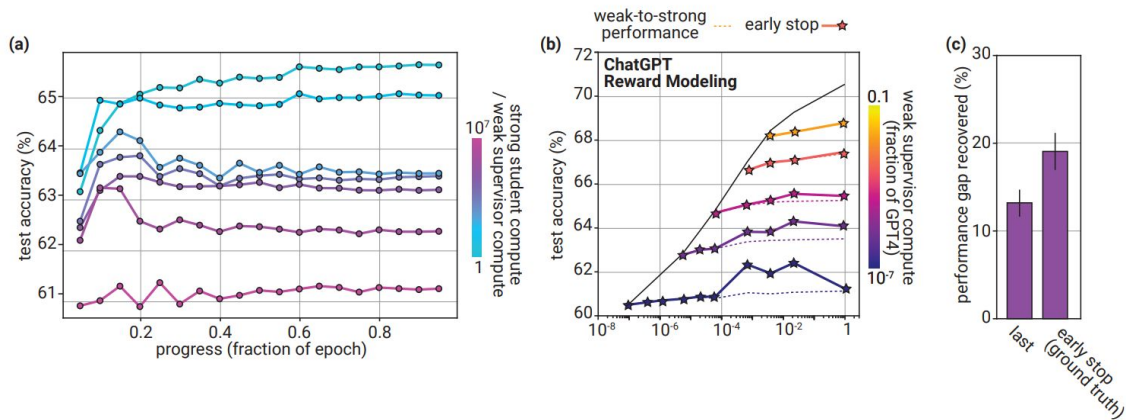


## Результаты

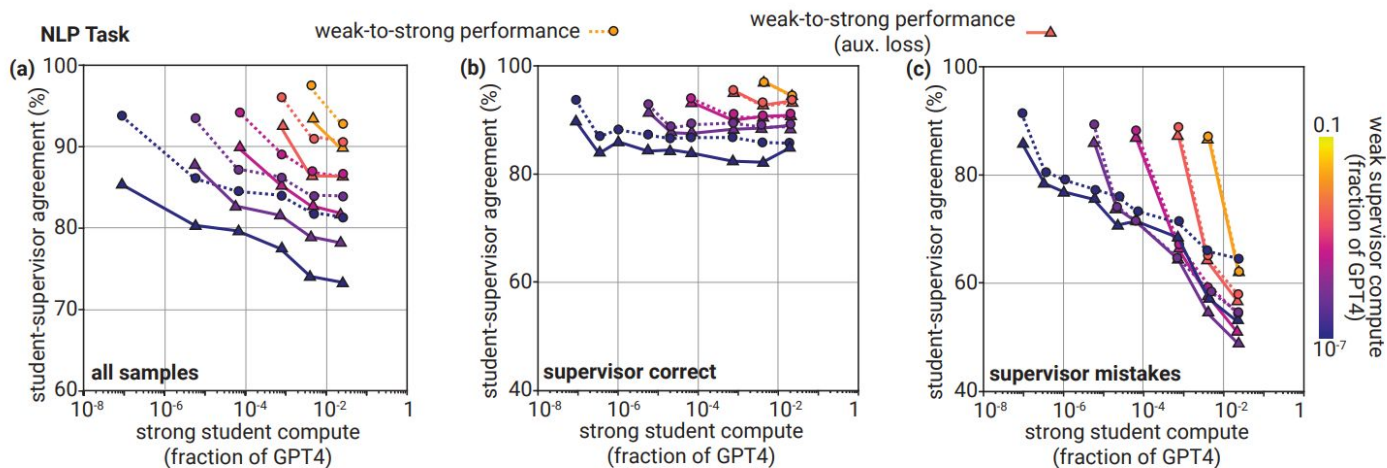
- Сильные модели-ученики все-таки обучаются с помощью слабых учителей
- Наивный метод - есть рост PGR
- Бутстрэп - улучшение PGR, особенно для шахмат
- auxiliary loss еще улучшает качество, потому что модель-ученик начинает меньше копировать ошибки учителя

# Understanding weak-to-strong generalization

- Копирование учителя

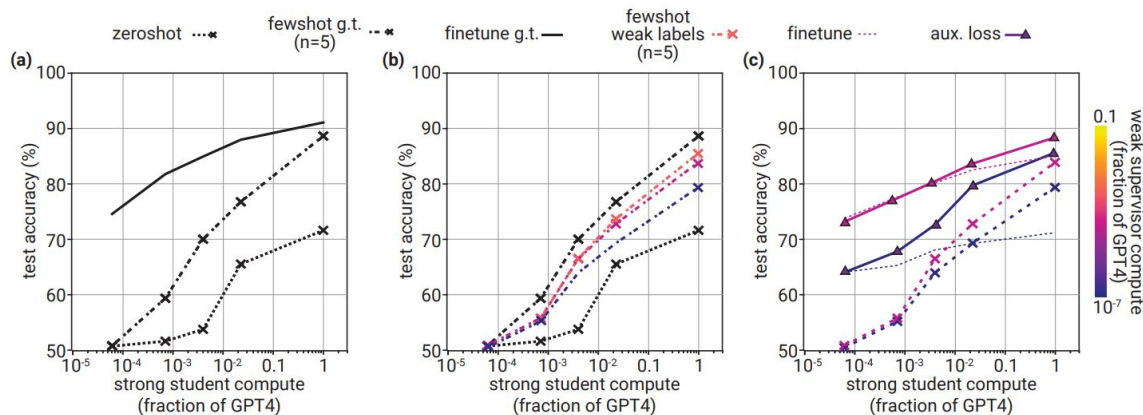


# Student-supervisor agreement



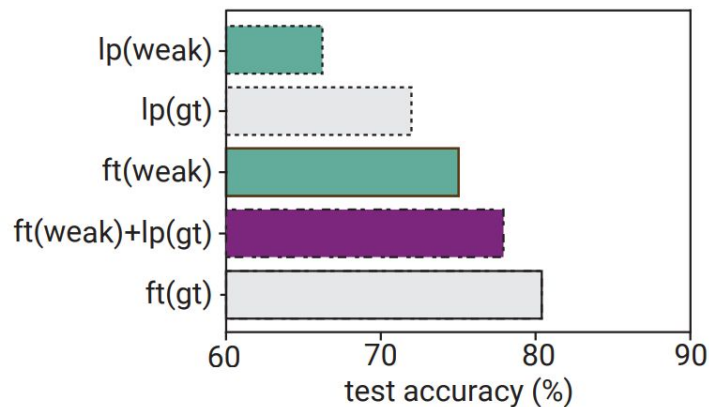
# Understanding weak-to-strong generalization

- Легко ли ученику “вспомнить”?



# Understanding weak-to-strong generalization

- Хорошо ли задача интерпретируется?





## Итог

- Сильные модели-ученики все-таки обучаются с помощью слабых учителей
- Наивный метод - есть рост PGR
- Бутстрэп - улучшение PGR, особенно для шахмат
- auxiliary loss еще улучшает качество, потому что модель-ученик начинает меньше копировать ошибки учителя
- **zero-shot и few-shot prompting на большой модели-ученике работает почти также, как weak-to-strong generalization**
- **weak-to-strong generalization делает задачу более интерпретируемой для модели**





## Сильные/слабые стороны статьи

- + Глубокий обзор поставленной проблемы
- + Понятный стиль повествования статьи
- Разобран небольшой круг задач и моделей



## Использованная литература

- <https://cdn.openai.com/papers/weak-to-strong-generalization.pdf>