Beyond RLHF

CSE 5539: Advanced Topics in Natural Language Processing

https://shocheen.github.io/courses/advanced-nlp-fall-2024

Logistics

1. Optional Self-review: Assignment up on Canvas
1. Mid-way report: Due November 4

Today's goal

Previously we studied Reinforcement Learning with Human Feedback. Today, we will look at works which question:

1. Is Reinforcement Learning needed for align to human preferences?
2. Are humans capable of provided preferences all the time? What do we do if not?
    a. The paper is interesting foray into "synthetic data generation"

Part I: No RL

Direct Preference Optimization:

Your Language Model is Secretly a Reward Model

Outline

1. Motivation of Problem
2. RLHF Overview
3. DPO Intuition
4. DPO in action

DPO

make **learning** from **preferences** easier by

avoiding

Reward Models and **Reinforcement Learning**

is a banana a fruit or a herb?

A banana is a fruit.

A banana is actually both a fruit and an herb. In botanical terms, the banana is a fruit because it contains the seeds of the plant, even though…

Typical RLHF

1. **SFT**
2. **Reward Modeling**
3. **RL Fine-Tuning**

Typical RLHF

1. **SFT**
   a. Start with model fine-tuned on high quality data from a downstream task
2. **Reward Modeling**
3. **RL Fine-Tuning**

Typical RLHF

1. **SFT**
   a. Start with model fine-tuned on high quality data from a downstream task
2. **Reward Modeling**
   a. SFT model is prompted for multiple responses to a query
   b. Humans rank the responses
   c. Train a (proxy) reward model to differentiate responses
3. **RL Fine-Tuning**

Typical RLHF

1. **SFT**
   a. Start with model fine-tuned on high quality data from a downstream task
2. **Reward Modeling**
   a. SFT model is prompted for multiple responses to a query
   b. Humans rank the responses
   c. Train a (proxy) reward model to differentiate responses
3. **RL Fine-Tuning**
   a. Online: gather training samples after each learning update
   b. Use PPO to update optimal policy using scores from reward model **(2)**

✍️: Patrick Da Silva

DPO Motivation

Benefits from avoiding steps 2 and 3

Hardware: no reward model
Efficiency: no online sampling
Stability: no PPO hyperparameters

1. **SFT**
   a. Start with model fine-tuned on high quality data from a downstream task
2. **~~Reward Modeling~~**
   a. ~~SFT model is prompted for multiple responses to a query~~
   b. ~~Humans rank the responses~~
   c. ~~Train a (proxy) reward model to differentiate responses~~
3. **~~RL Fine-Tuning~~**
   a. ~~Online: gather training samples after each learning update~~
   b. ~~Use PPO to update optimal policy using scores from reward model (2)~~

✍: Patrick Da Silva

DPO Intuition

Question

How to create a loss function that derives an

OPTIMAL POLICY DIRECTLY from rewards?

✍️: Patrick Da Silva

DPO Intuition

Question

How to create a loss function that derives an

OPTIMAL POLICY DIRECTLY from rewards?

Answer

1. **Reparameterize the Bradley Terry Model**
2. **Transform loss over reward functions into a loss function over policies**

✍: Patrick Da Silva

DPO Intuition: Reparametrize Bradley Terry

describes human preference distribution **p***

As a function of **reward (RLHF)**

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(\boxed{r^*(x, y_1)}\right) + \exp\left(r^*(x, y_2)\right)}.$$

As a function of **policy (DPO)**

$$p^*(y_1 \succ y_2 \mid x) = \frac{1}{1 + \exp\left(\boxed{\beta \log \frac{\pi^*(y_2|x)}{\pi_{\text{ref}}(y_2|x)}} - \beta \log \frac{\pi^*(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right)}$$

π* is the optimal policy   **π_{ref}** is the initialized policy

✍: Patrick Da Silva

DPO Intuition: Transform Loss

Parameterized reward function **(RLHF)**

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma(\boxed{r_\phi(x, y_w)} - r_\phi(x, y_l))\right]$$

Parameterized policy **(DPO)**

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\left(\boxed{\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

Experimental Validations: Evaluation tasks

- **Controlled sentiment generation**
    - Given a prefix x from the IMDb dataset, policy produces y with positive sentiment
- **Summarization**
    - Reddit TL;DR Dataset
- **Single-turn dialogue**
    - Anthropic Helpful and Harmless dialogue dataset

Experimental Validations: Models/Methods

1. **Preferred-FT:** Pythia-2.8B trained on y_w
2. **Unlikelihood:** maximize the probability assigned to y_w and minimize the probability assigned to y_l
3. **PPO**: trained from preference data
4. **PPO-GT:** trained from ground-truth RM in controlled sentiment generation
5. **Best of N:** sample n responses and return the highest scoring response according to a RM learned from the preference data
6. **DPO**

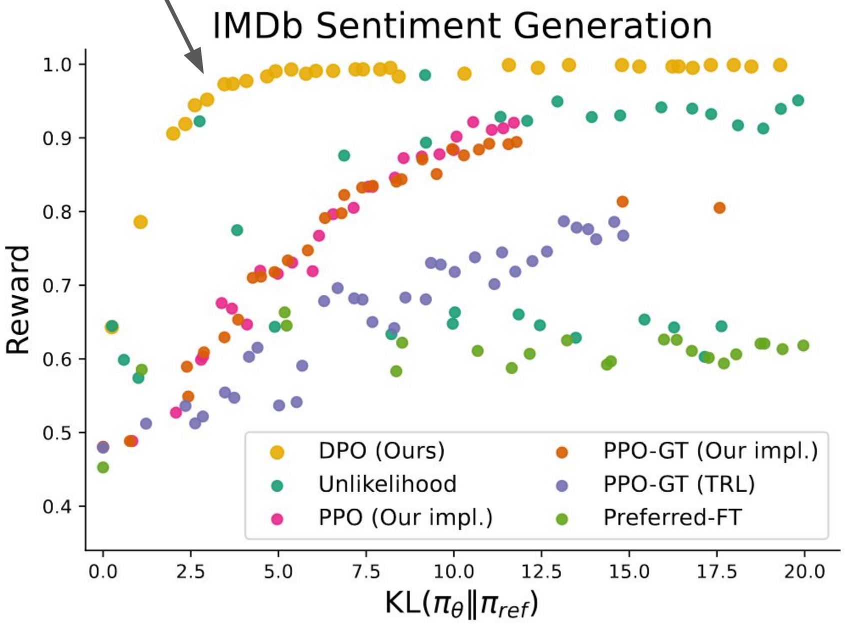
✍: Patrick Da Silva

Experimental Validations: Basic Objective

KL Divergence: How far has the optimal policy moved from the initial model

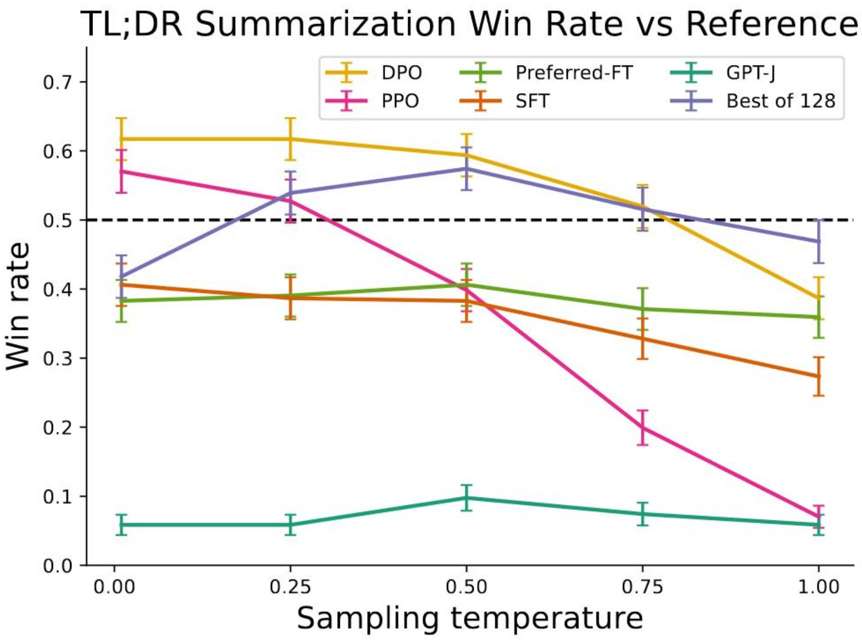
(Notice early peak, more efficient)

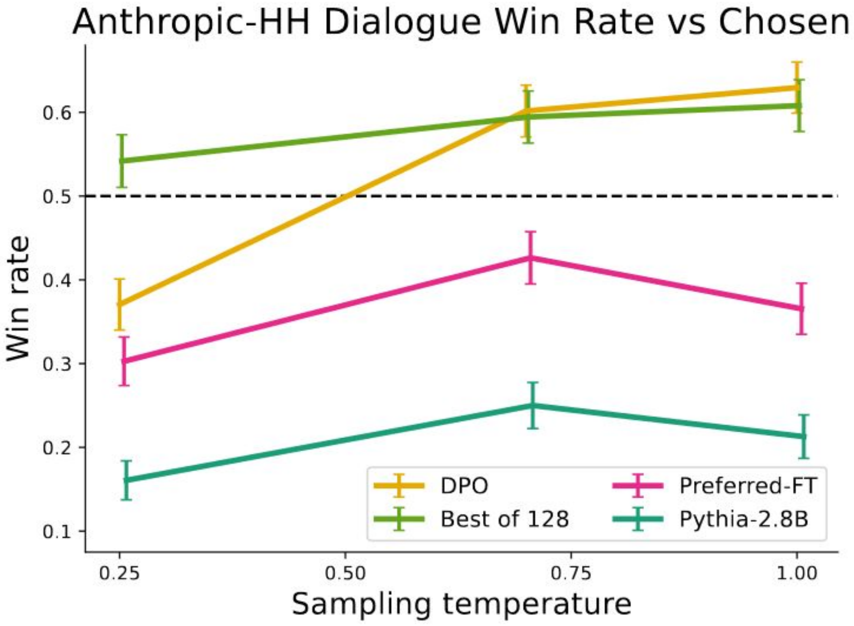
Large KL divergence is not desirable



✍: Patrick Da Silva

Experimental Validations: “Hard” Tasks





✍: Patrick Da Silva

Experimental Validations: OOD Generalization

Switch Tasks:

- News summarization rather than Reddit TL;DR

1. **DPO** outperforms **PPO**
2. **Initial evidence** that **DPO** can generalize as well as **PPO**

	Win rate vs. ground truth	
Alg.	Temp 0	Temp 0.25
DPO	0.36	0.31
PPO	0.26	0.23

Table 1: GPT-4 win rates vs. ground truth summaries for out-of-distribution CNN/DailyMail input articles.

Direct Preference Optimization: Your Language Model is Secretly a Reward Model -

Scientific Reviewer

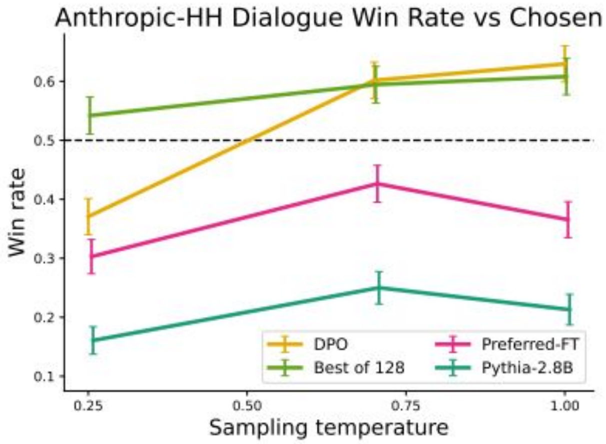
Bowei Kou

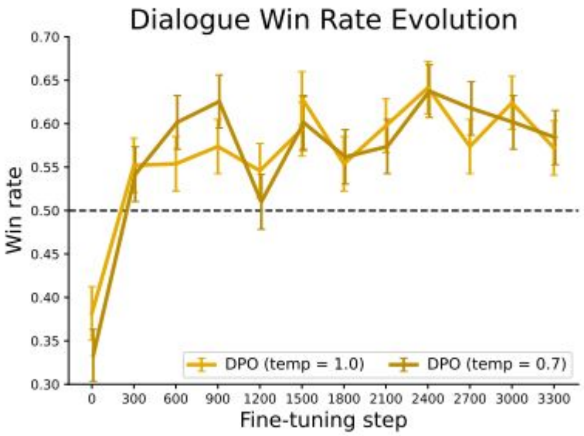
Strengths

Creative idea: This paper presents a new perspective. Treating LM as an RM, thereby reducing complexity in the optimization process while maintaining alignment with human preferences.

Clear-cut theory: The paper provides a solid theoretical foundation for DPO and clearly explains how DPO works. i.e. linking the softmax transform.

Sufficient results: The paper provides sufficient experimental results to demonstrate that DPO performs well in several tasks and is able to compete with current method.





	DPO	SFT	PPO-1
N respondents	272	122	199
GPT-4 (S) win %	47	27	13
GPT-4 (C) win %	54	32	12
Human win %	58	43	17
GPT-4 (S)-H agree	70	77	86
GPT-4 (C)-H agree	67	79	85
H-H agree	65	-	87

Table 2: Comparing human and GPT-4 win rates and per-judgment agreement on TL;DR summarization samples. **Humans agree with GPT-4 about as much as they agree with each other.** Each experiment compares a summary from the stated method with a summary from PPO with temperature 0.

Weakness

Potential risks of overfitting: Direct optimization of preferences may increase the risk of overfitting to the training dataset, especially if the preference data does not represent a wide range of people well.

Data quality: Direct preference optimization relies on high-quality preference data, and there is insufficient discussion in the paper on dealing with noise or inconsistency in preference data, which may lead to optimization failures in real-world applications

Review

- Novelty	4.0/5
- Correctness	4.5/5
- Clarity	4.0/5
- Significance	4.0/5
- Recommendation:	Accept

DPO - ARCHAEOLOGIST 🏺

Abraham Owodunni

Main Motivation for DPO

Prior work: Alignment with RLHF is slow, complex and expensive,

DPO: what can we do about this?

How it started:

It all started in 1952:

Bradley-Tary Model: *Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons*

How it started:

It all started in 1952:

Bradley-Tary Model: *Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons*

Idea: given n independent sample (a, b,c … n), how can we say *a* is better than *b* if you pair the samples?

How it started:

It all started in 1952:

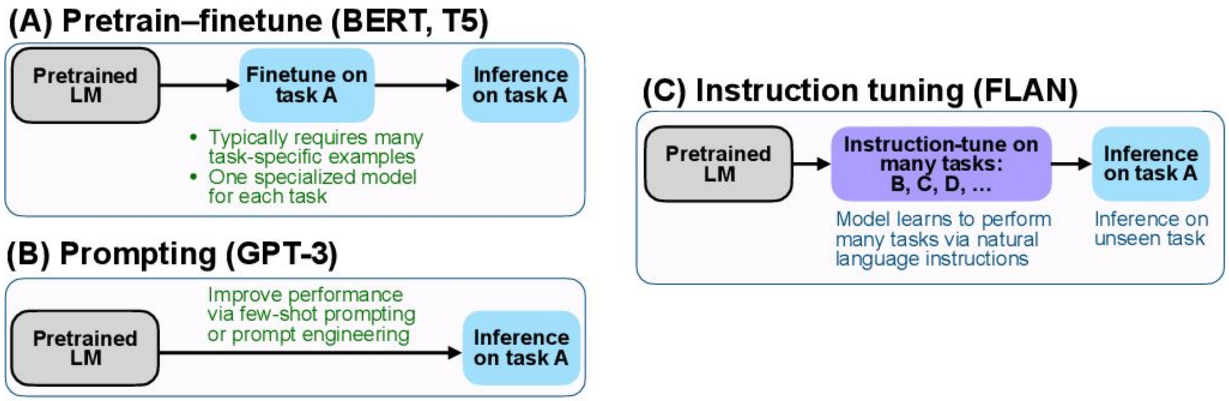
Bradley-Tary Model: *Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons*

Idea: given n independent sample (a, b,c … n), how can we say a is better than b if you pair the samples?

$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}.$$

How it started:

- People started Instruction-tuning
    - Wei et al., (2021) Finetuned language models are zero-shot learners. - (from previous class)



How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human preference
    -
-

How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human preference
    - Summarization: Ziegler et al., 2020 (OpenAI). Fine-Tuning Language Models from Human Preferences
-

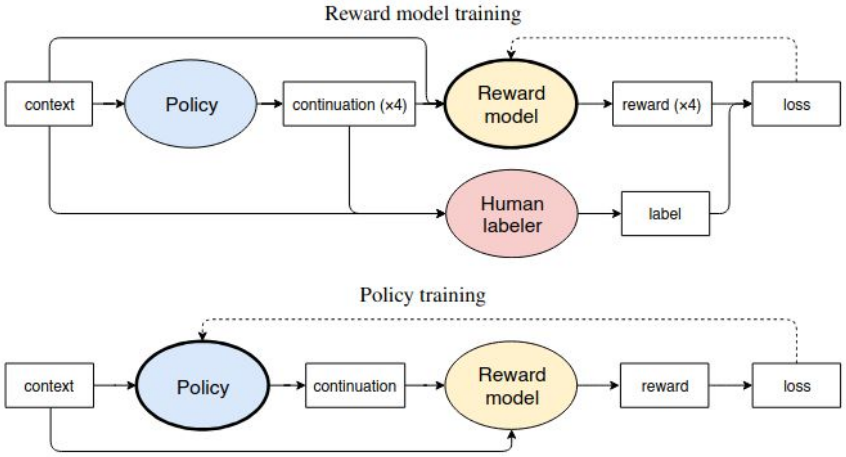
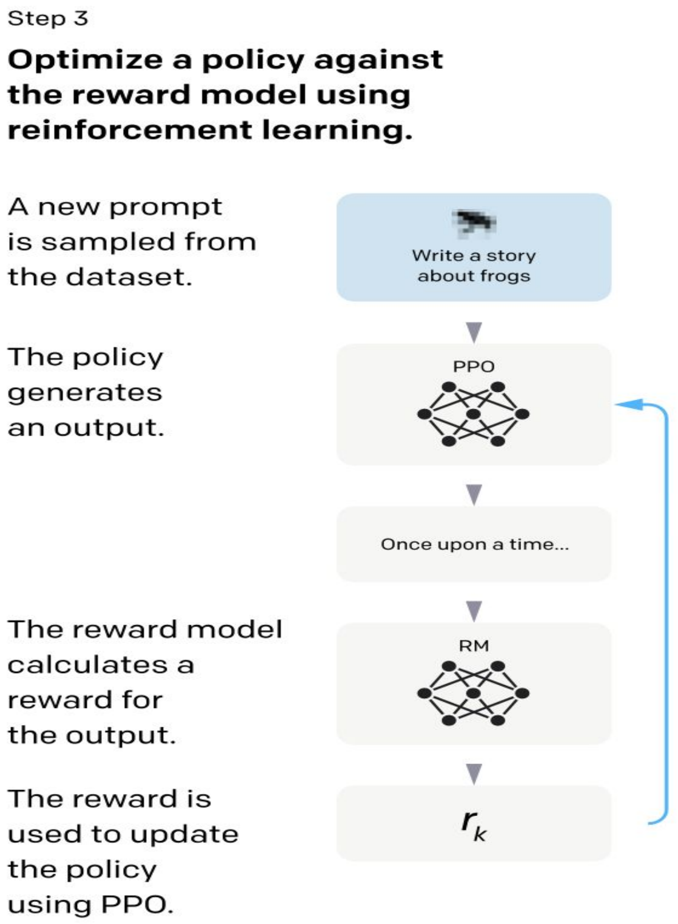


Figure 1: Our training processes for reward model and policy. In the online case, the processes are interleaved.

How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human
    - Summarization: Ziegler et al., 2020 (OpenAI). Fine Preferences.
- And Lastly, alignment with RLHF:
    - OpenAI, 2022: Training language models to follow Instruction with human feedback.



How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human preference
    - Summarization: Ziegler et al., 2020 (OpenAI). Fine-Tuning Language Models from Human Preferences.
- And lastly, alignment with RLHF: But RLFH is slow, complex and expensive.

How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human preference
    - Summarization: Ziegler et al., 2020 (OpenAI). Fine-Tuning Language Models from Human Preferences.
- And lastly, alignment with RLHF: But RLFH is slow, complex and complicated.
- DPO (2024): You can simply switch from:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left[r_\phi(x,y)\right] - \beta \mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(y \mid x) \,||\, \pi_{\mathrm{ref}}(y \mid x)\right],$$

How it started:

- People started Instruction-tuning
- They moved to Instruction-tuning on human preference
    - Summarization: Ziegler et al., 2020 (OpenAI). Fine-Tuning Language Models from Human Preferences.
- And lastly, alignment with RLHF: But RLFH is slow, complex and complicated.
- DPO (2024): You can simply switch from:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[r_\phi(x,y) \right] - \beta \mathbb{D}_{\mathrm{KL}} \left[\pi_\theta(y \mid x) \mid\mid \pi_{\mathrm{ref}}(y \mid x) \right],$$

to

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)} \right) \right]$$

Direct Preference Optimization:Your Language Model is Secretly a Reward Model

Visionary🔭

Jiachen Jiang

Extend DPO into the multimodal domain

- Given the success of Direct Preference Optimization (DPO) in replacing reinforcement learning for aligning models with human preferences efficiently, the follow-up project could extend DPO into the **multimodal domain**, unlocking new applications beyond text.

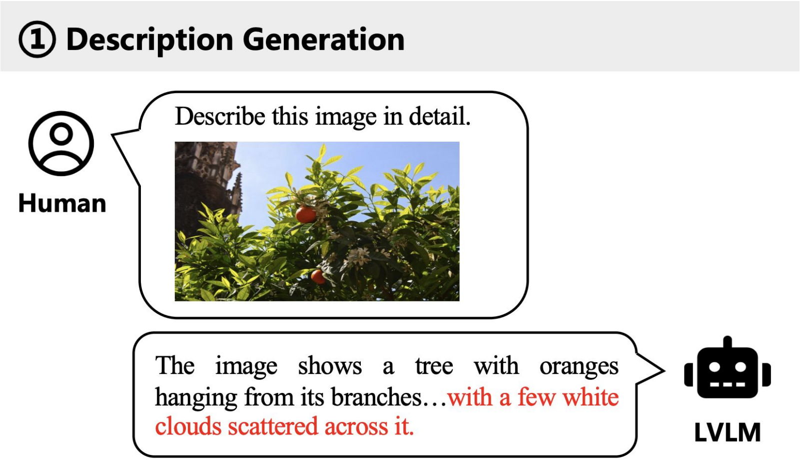
- The goal is to align the generation of multimodal outputs (e.g., text-based image captions, video summaries) with user preferences without relying on complex reinforcement learning pipelines.

✍: Jiachen Jiang

Research Questions

- **Cross-modal preference integration(Input):** How can human preferences for multiple types of outputs (text, audio, image) be effectively combined?
- **Multi-modal data alignment(Output):** Can the DPO framework efficiently optimize large models generating diverse outputs like descriptive captions, summaries, and instructions?

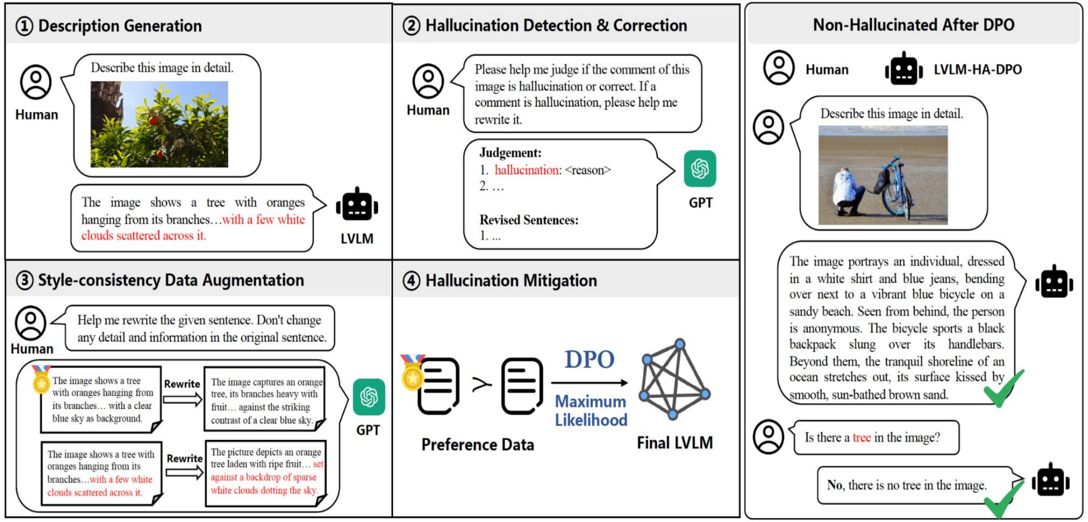
Hallucinations: models generate textual descriptions that inaccurately depict or entirely fabricate content from associated images



Existing research on this direction

HA-DPO(https://opendatalab.github.io/HA-DPO/) HA-DPO is designed to mitigate hallucinations in multimodal models

- The model is trained to favor the non-hallucinating response when presented with two responses of the same image (one accurate and one hallucinatory)
- It proposes an efficient pipeline for constructing positive (non-hallucinatory) and negative (hallucinatory) sample pairs, ensuring a highquality, style-consistent dataset for robust preference learning.
- Language models like GPT-4 are used to evaluate hallucination-free outputs
- HA-DPO has shown success in improving accuracy for models like MiniGPT-4, particularly in image-text alignment tasks



IO STATE UNIVERSITY

Part II: No HF

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

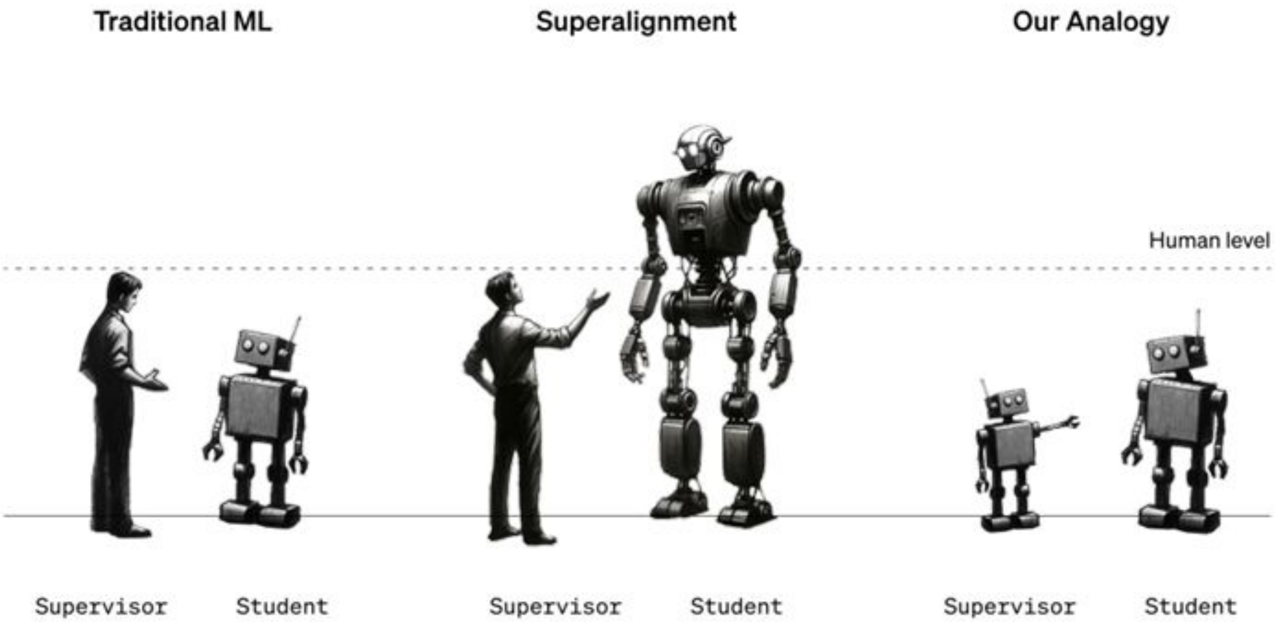
Stakeholder: Hanane Nour Moussa

The Superalignment Problem

- RLHF is the main method used to align today's models
  - Effective when human evaluators can understand the model behavior
- But what happens when humans try to align superhuman models?
- How do we ensure AI systems much smarter than humans follow human intent?

The Superalignment problem

- How can we study this problem today? We can consider the analogy of weak models supervising strong models



✍: Hanane Nour Moussa

The Superalignment Problem

- The setup: Finetuning **large** (aka, **strong**) pretrained models on labels generated by **small** (aka, **weak**) models and observing how they generalize.
- Two possibilities: **Imitation** or **Elicitation**
- The hypothesis: the strong model can generalize beyond the weak supervision and solve hard problems for which the weak supervisor can only give incomplete/flawed training labels ⇒ **Weak-to-strong generalization**

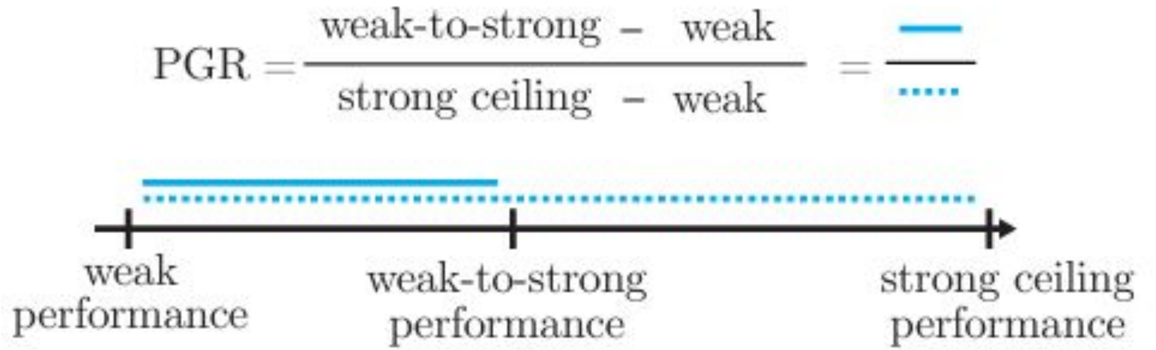
Methodology

For three types of tasks (NLP benchmarks, chess puzzles dataset, and internal ChatGPT reward modeling dataset), the authors:

- Create a **weak supervisor**: finetune small pretrained models on GT labels and use them to generate weak labels ⇒ **weak performance**
- Train a **strong student** model with weak supervision: finetune large models from the GPT-4 family spanning 7 orders of magnitude with the weak labels ⇒ **weak-to-strong performance**
- Train a strong model with GT labels as **ceiling**: finetune strong model with GT labels ⇒ **strong ceiling performance**

Methodology

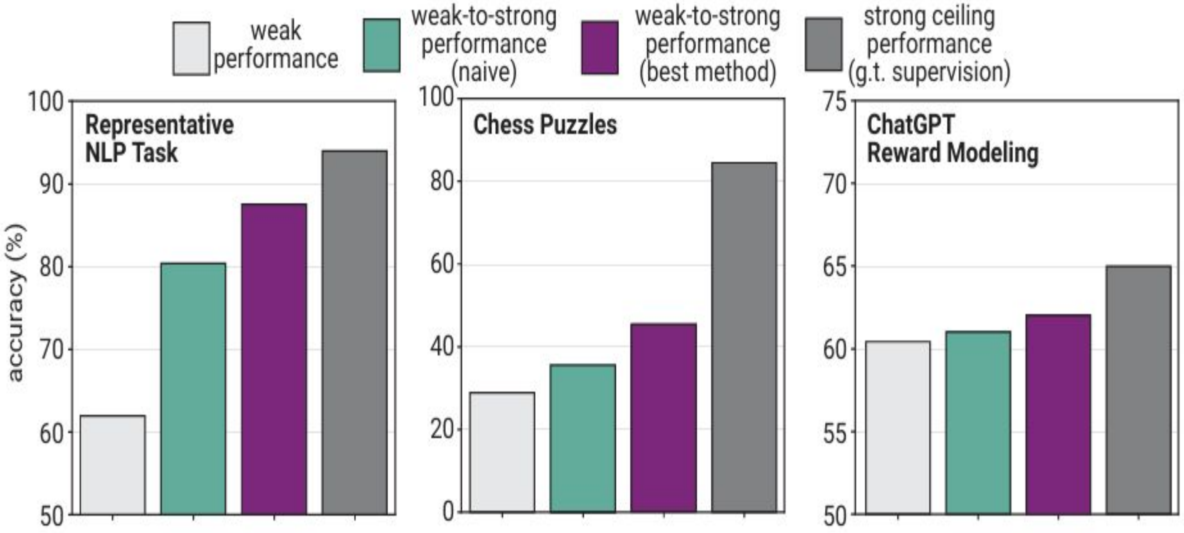
- Metric: **Performance Gap Recovered** (**PGR**). 0 <= PGR <= 1



✍️: Hanane Nour Moussa

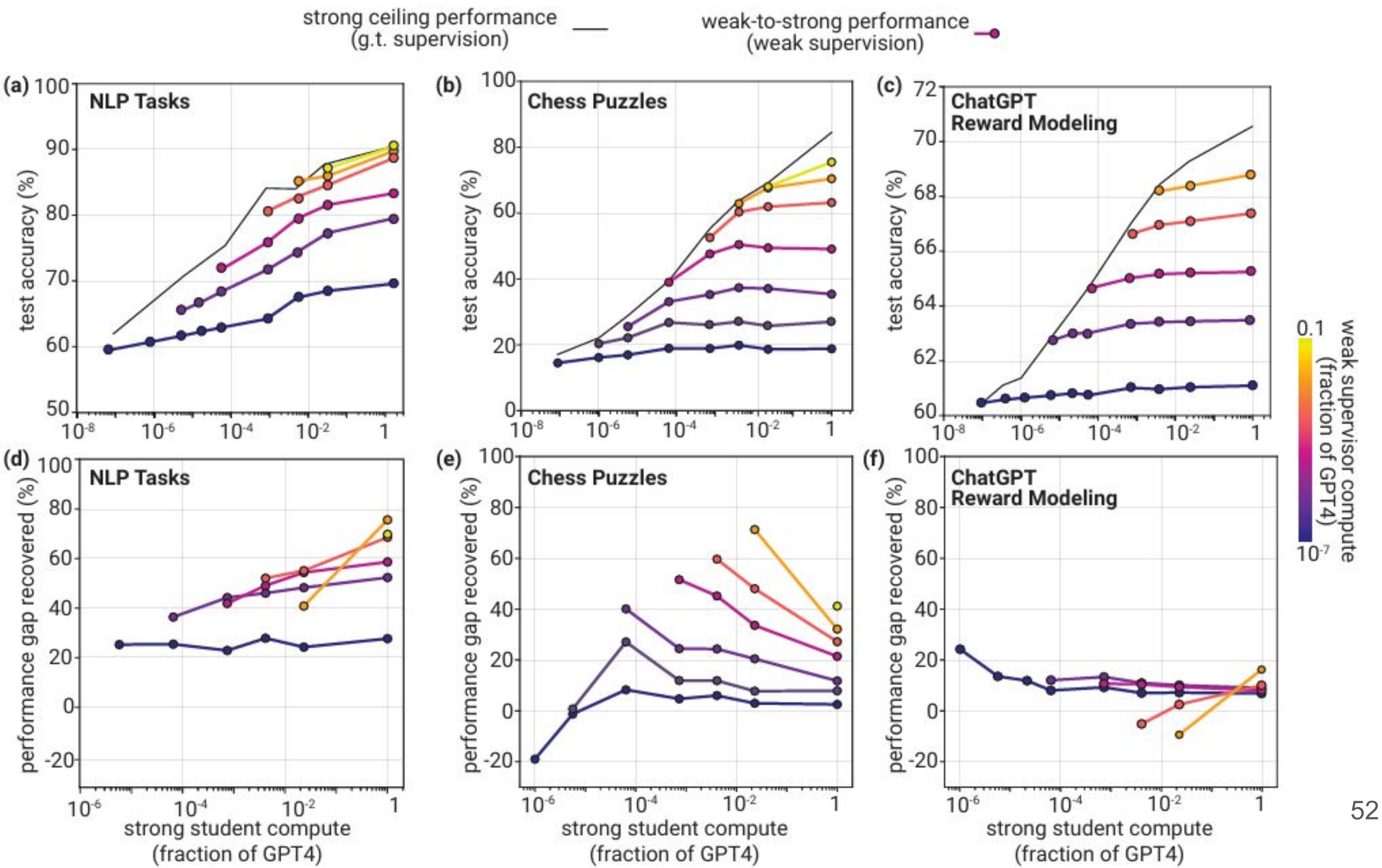
Main results

- Strong pretrained models naturally generalize beyond their weak supervisors
- Naively finetuning on weak supervision is not enough
- Improving weak-to-strong supervision is tractable



Results: Naive Finetuning on Weak Labels

Promising weak to strong generalization on NLP and chess, but poor performance on reward modeling



✍: Hanane Nour Moussa

Results: Naive Finetuning on Weak Labels

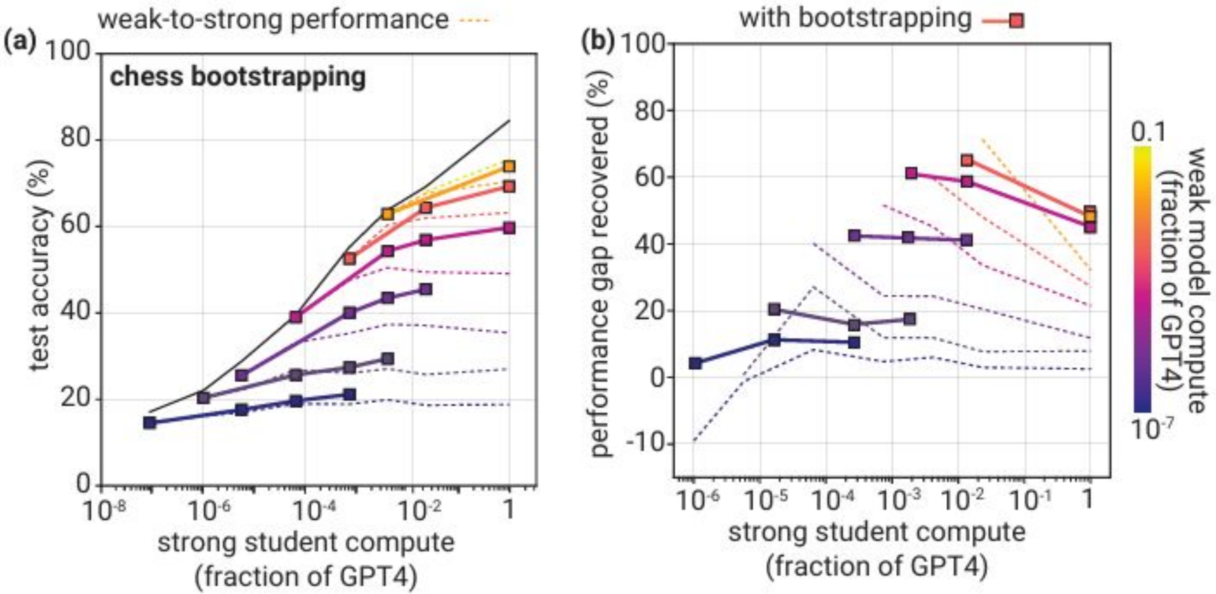
- In general, across all settings, weak-to-strong generalization holds true: **Strong students consistently outperform their weak supervisors**
- Two conclusions to make:
    - Weak-to-strong learning is a tractable problem
    - Naive weak, human level supervision will be insufficient to align strong, superhuman models
- How can we improve weak-to-strong generalization?

Improving weak-to-strong generalization

- Two approaches offer proofs-of-concept:
  - Bootstrapping with intermediate model sizes
  - Auxiliary Confidence Loss

Bootstrapping with intermediate model sizes

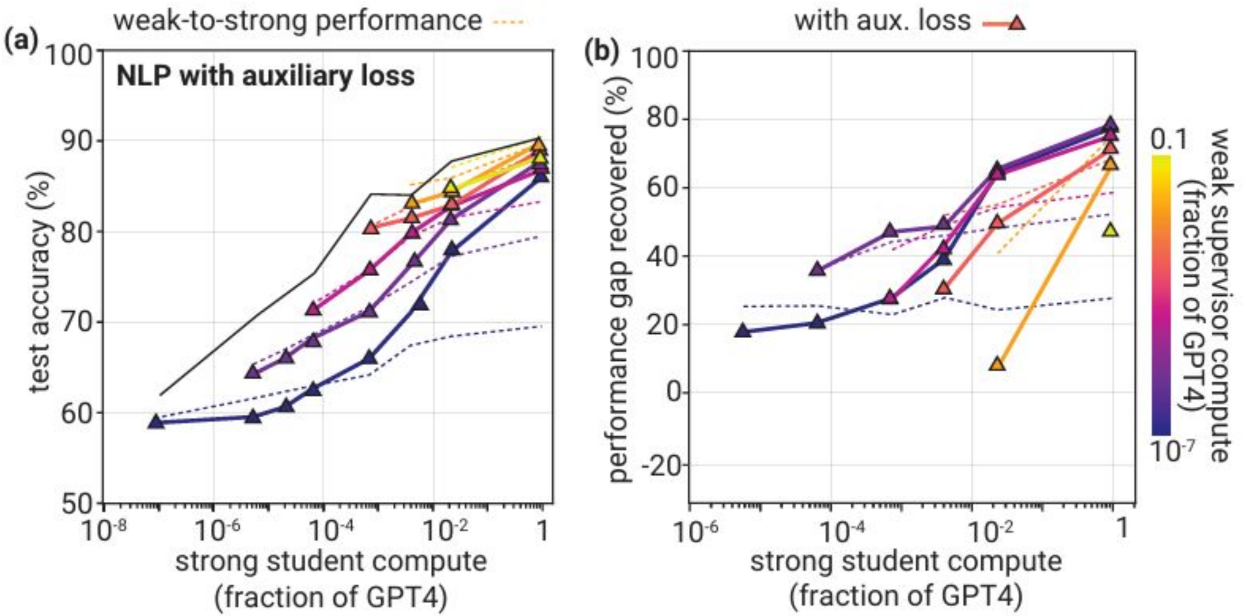
- Idea: Construct a sequence of models M_1 → M_2 → … → M_n of increasing sizes. Use weak labels from M_i to finetune M_i+1. Improves performance in chess setting.



✍️: Hanane Nour Moussa

Auxiliary Confidence Loss

- Idea: Adding an auxiliary confidence loss term to the standard cross entropy objective. This reinforces the strong model's confidence in its own predictions even when they disagree with the weak labels (learn intent, not errors)



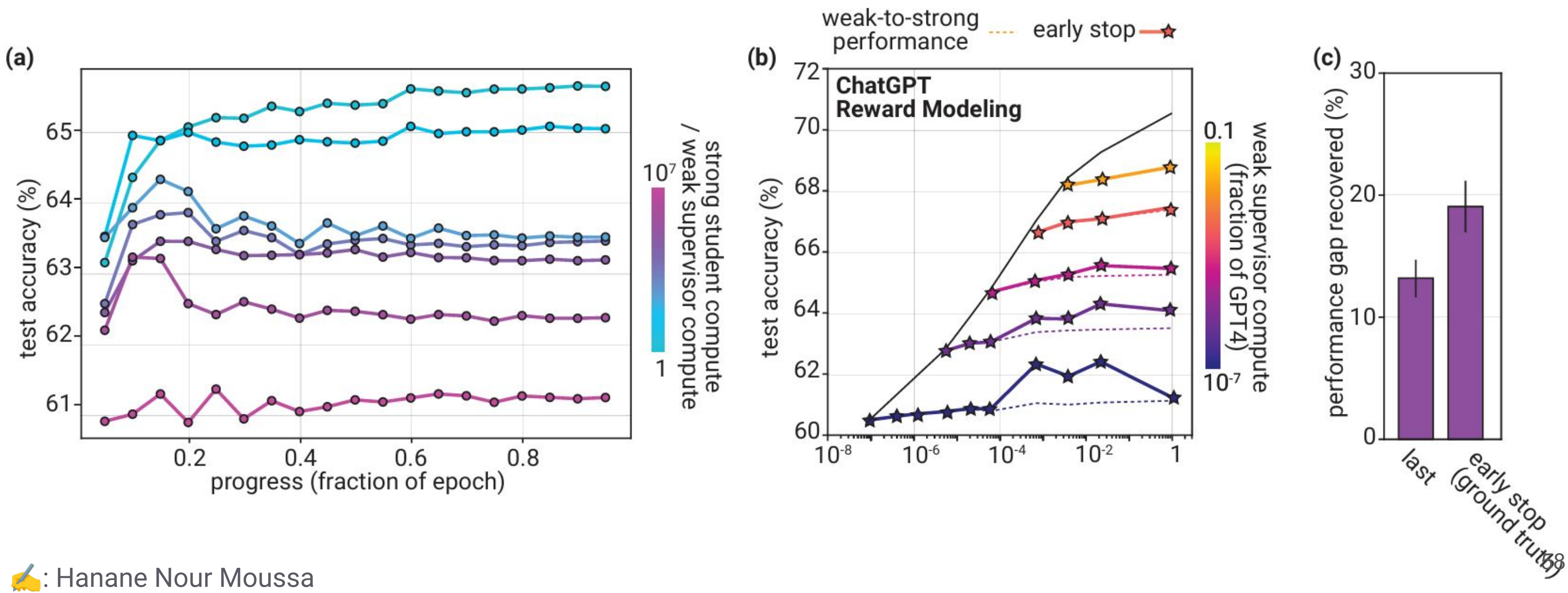
✍️: Hanane Nour Moussa

Understanding Weak-to-Strong Generalization

- In order to develop effective methods for solving superalignment, we need to understand **when and why** they work.
- Two phenomena are investigated:
  - Imitation of supervisor mistakes
  - Salience of the tasks to the strong student model

Understanding Imitation

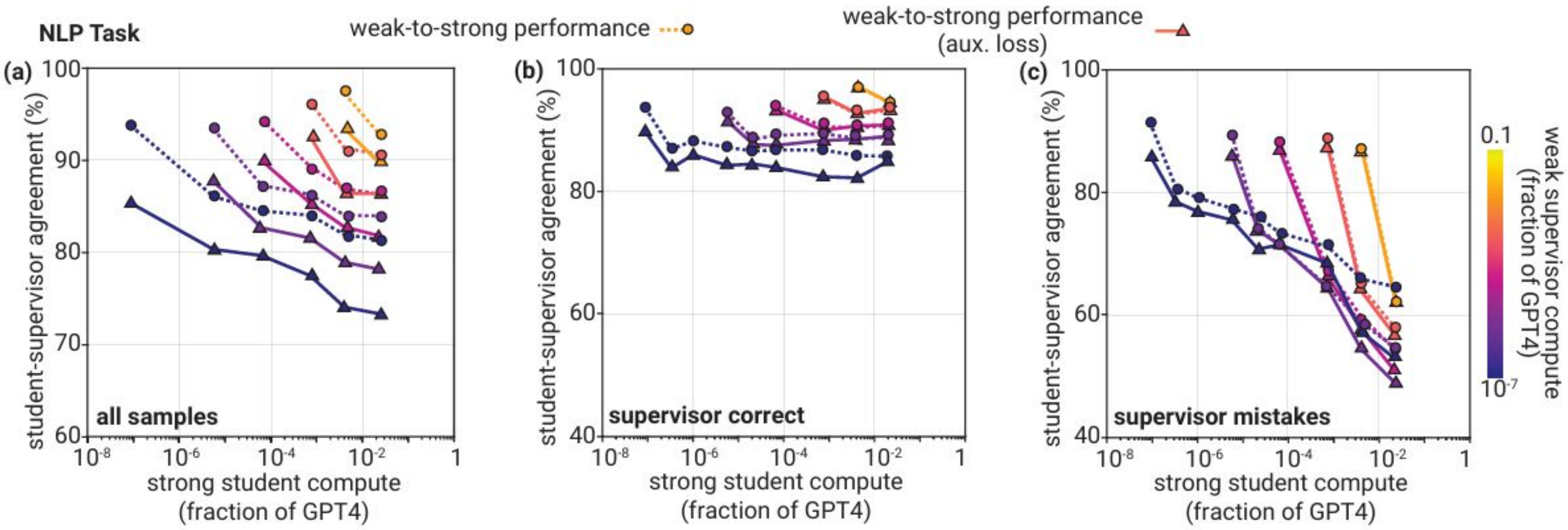
- Overfitting to weak supervision: Strong models overfit to weak labels



✍: Hanane Nour Moussa

Understanding Imitation

- Student-supervisor agreement is reduced with auxiliary loss



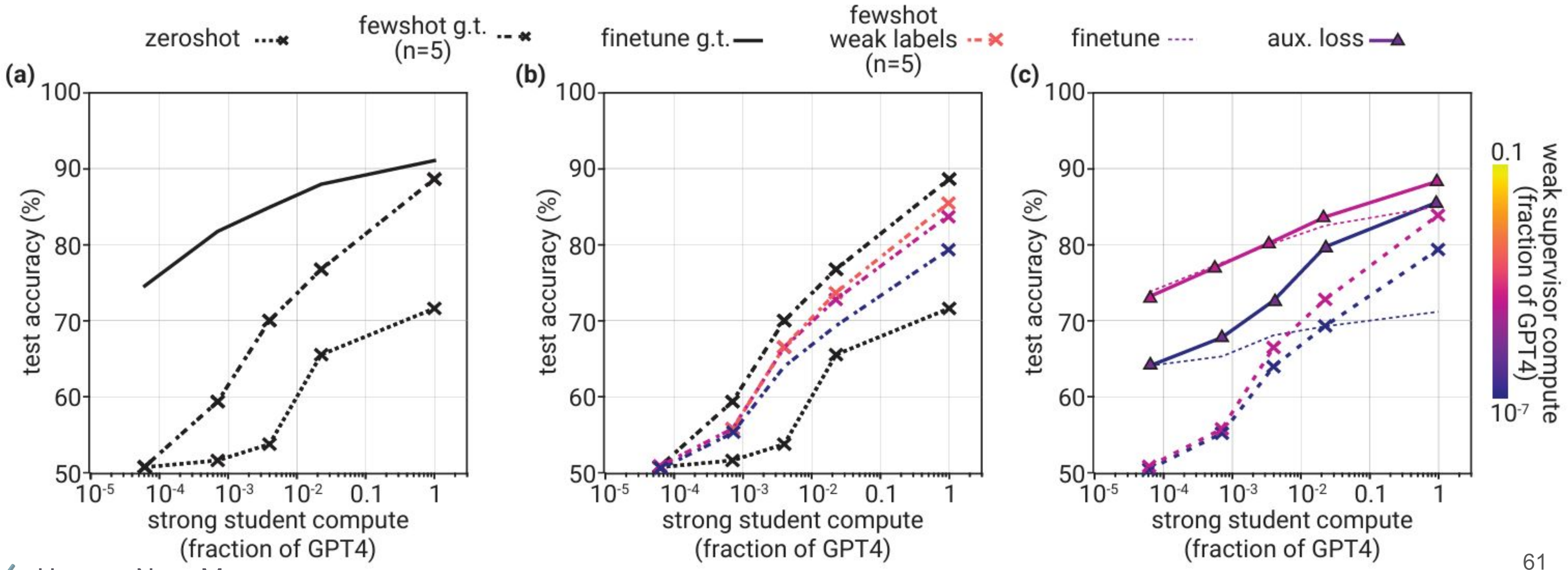
✍️: Hanane Nour Moussa

Saliency in the strong model representations

- Weak-to-strong generalization might be particularly feasible when the task we want to elicit is internally "**salient**" to the strong model.

Saliency in the strong model representations

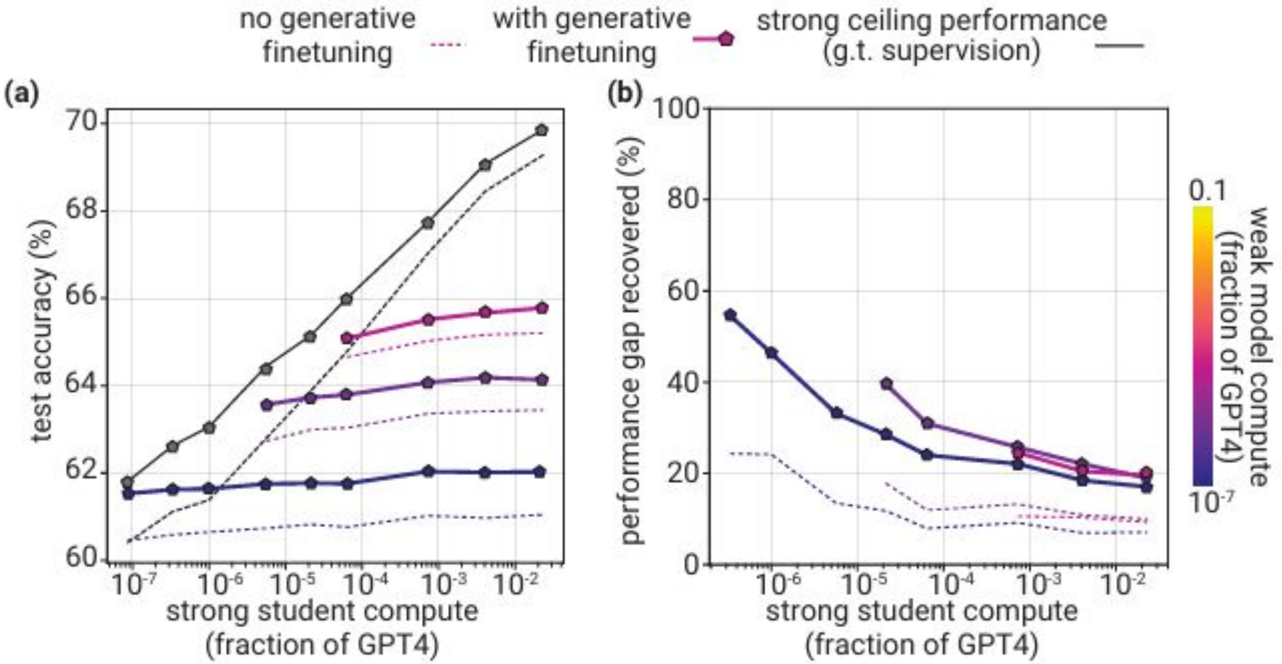
- Eliciting strong model knowledge with prompting (results average across 7 NLP tasks). It’s relatively easy to elicit knowledge from larger student models.



✍: Hanane Nour Moussa

Saliency in the strong model representations

- Generative supervision (unsupervised finetuning) on reward modeling improves weak-to-strong performance and PGR



✍: Hanane Nour Moussa

Remaining Disanalogies

- **Imitation saliency**: superhuman models will be very good at predicting human behavior and may thus easily imitate weak human errors. This is not captured in the paper's experimental setup
- **Pretraining leakage**: superhuman knowledge models may be latent, not observable. Superhuman models may never directly observe superhuman alignment relevant capabilities. They will be predominantly "latent" and thus harder to elicit.

⇒ May cause results to be overly optimistic

Future work

- Analogous setups
  - Fixing disanalogies or validating that they are not severe
  - Adding more more complex generative tasks
  - Identifying new and more specific disanalogies

- Strong scientific understanding
  - A thorough understanding of when and why methods work
  - Why does naive finetuning work better for NLP tasks compared to reward modeling?
  - What makes a concept easy or hard to elicit? How can saliency be defined?

No HF

Reviewer

Junjie Zhang

Strengths

- Clarity

This cartoon clearly introduces the problem defined in this paper, allowing the reader to quickly understand the main point of the article.

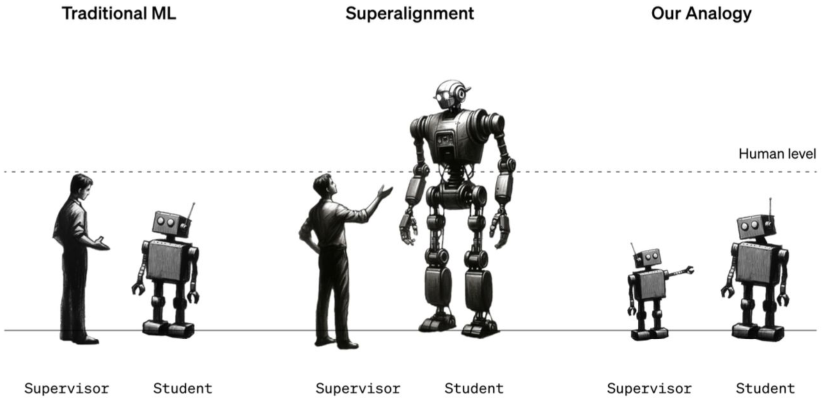


Figure 1: **An illustration of our methodology.** Traditional ML focuses on the setting where humans supervise models that are weaker than humans. For the ultimate superalignment problem, humans will have to supervise models much smarter than them. We study an analogous problem today: using weak models to supervise strong models.

Strengths

For example, if a model can generate complicated code, then it should intuitively also know whether that code faithfully adheres to the user's instructions. As a result, for the purposes of alignment we do not need the weak supervisor to teach the strong model new capabilities; instead, we simply need the weak supervisor to elicit what the strong model *already knows*.

1. **Create the weak supervisor.** Throughout most of this work, we create weak supervisors by finetuning small pretrained models on ground truth labels.We call the performance of the weak supervisor the *weak performance*, and we generate *weak labels* by taking the weak model's predictions on a held-out set of examples.

2. **Train a strong student model with weak supervision.** We finetune a strong model with the generated weak labels. We call this model the *strong student model* and its resulting performance the *weak-to-strong performance.*

3. **Train a strong model with ground truth labels as a ceiling.** Finally, for comparison, we finetune a strong model with ground truth labels.[4] We call this model's resulting perfor- mance the *strong ceiling performance*. Intuitively, this should correspond to "everything the strong model knows," i.e. the strong model applying its full capabilities to the task.

Strengths

- Quality

The methodologies are well-explained. The authors did detailed tests on three datasets, including varying the size of student model size and supervisor model size.

- Originality

This paper focuses on the alignment issues of future superhuman models, is highly original. The concept of weak-to-strong generalization is a novel and important contribution to the field of model alignment.

Strengths

- Soundness

The soundness of the paper is solid, with detailed empirical studies supporting the assumptions. The experiments demonstrate consistent outcomes across various tasks and model sizes.

- Broader Impact

The broader impact of this research is significant. The study aim to addresses the future challenges of AI alignment, which is crucial for developing safe and reliable AI systems.

Weaknesses

- limited tasks
- Still has a significant gap compared to the strongest student models
- Pretraining leakage
- Imitation saliency
- It is currently just a proof of concept and cannot be deployed on existing models.

Review

- Novelty 10/10 : This research focuses on the security issues of future super models, which makes it highly novel.
- Correctness 8/10 : Since super models have not yet emerged, some potential issues cannot be validated.
- Clarity 9/10 : The hypothesis is clearly stated and supported by reasonable experimental validation.
- Significance 10/10 : This research is important for AI safety.
- Recommendation Accept

Weak to Strong Generalization Archaeologist

Suchit Gupte

The Ohio State University

What inspired this work?

1. Snorkel
2. Self-training
3. Mean teachers are better role models
4. DivideMix

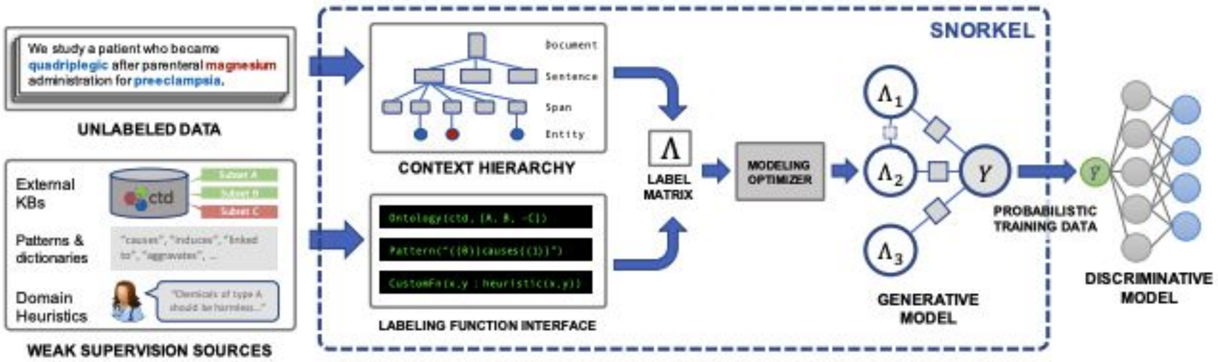
Snorkel

Snorkel: Rapid Training Data Creation with Weak Supervision

Alexander Ratner Stephen H. Bach Henry Ehrenberg
Jason Fries Sen Wu Christopher Ré
Stanford University
Stanford, CA, USA
{ajratner, bach, henryre, jfries, senwu, chrismre}@cs.stanford.edu

- Snorkel allows users to generate weak labels programmatically using labeling functions, reducing manual data annotation.

Relevance:
Provides a framework for combining weak supervision sources - Training models in low-label environments



: Suchit Gupte

The Ohio State University

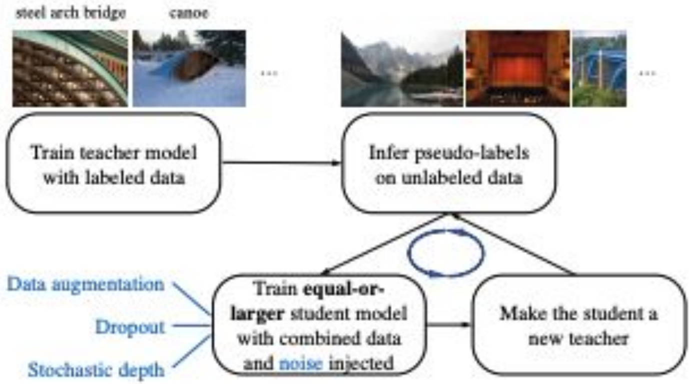
Self-training

Self-training with Noisy Student improves ImageNet classification

Qizhe Xie*[1], Minh-Thang Luong[1], Eduard Hovy[2], Quoc V. Le[1]
[1]Google Research, Brain Team, [2]Carnegie Mellon University
{qizhex, thangluong, qvl}@google.com, hovy@cmu.edu

- Combines teacher-student training where the student is trained with noise added to input and model parameters using pseudo-labels generated by the teacher.

Relevance:
Shows that using weak supervision along with noise regularization improves generalization and makes the model more robust.



Mean teachers are better role models

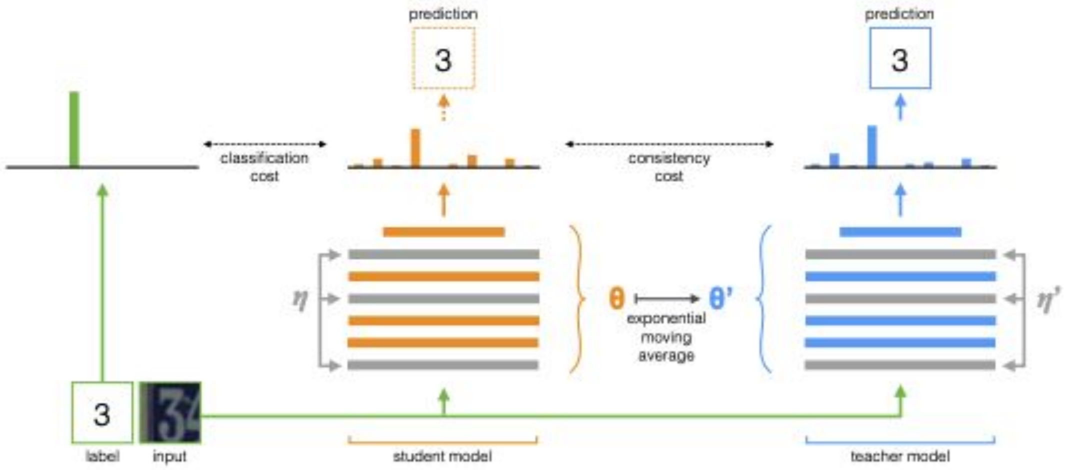
Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results

Antti Tarvainen
The Curious AI Company
and Aalto University
antti.tarvainen@aalto.fi

Harri Valpola
The Curious AI Company
harri@cai.fi

- The teacher model's parameters are an exponential moving average of the student model, encouraging consistent predictions between teacher and student on labeled and unlabeled data.

Relevance:
Highlights the power of consistency regularization in a weak supervision setting.



: Suchit Gupte

The Ohio State University

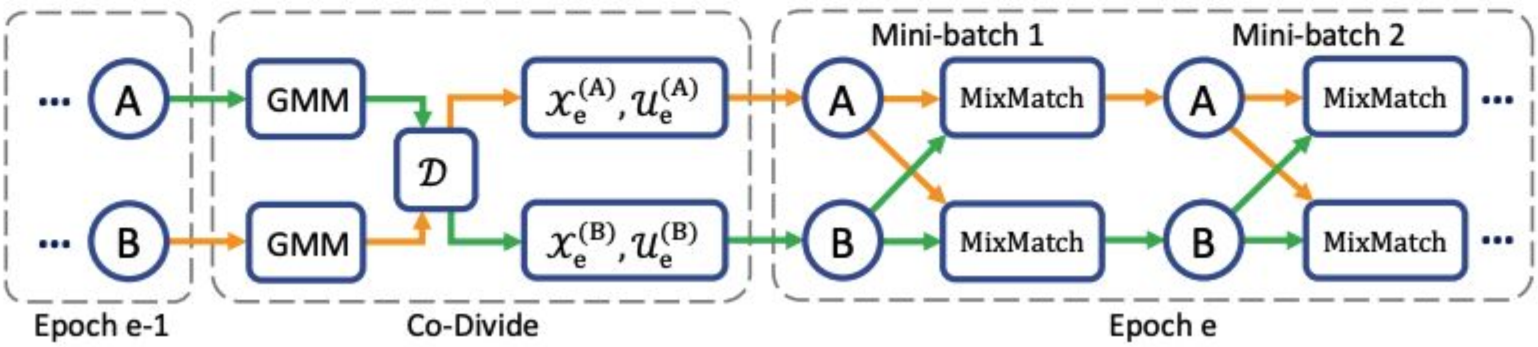
DivideMix

DIVIDEMIX: LEARNING WITH NOISY LABELS AS SEMI-SUPERVISED LEARNING

Junnan Li, Richard Socher, Steven C.H. Hoi
Salesforce Research
{junnan.li,rsocher,shoi}@salesforce.com

- Treats noisy labels as a form of weak supervision by framing the learning problem as semi-supervised learning.

Relevance:
Illustrates that noisy label problems can be approached with semi-supervised learning techniques, allowing models to generalize well despite label noise.



: Suchit Gupte

THE OHIO STATE UNIVERSITY

What this work inspired?

1. Self-Rewarding Language Models
2. Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

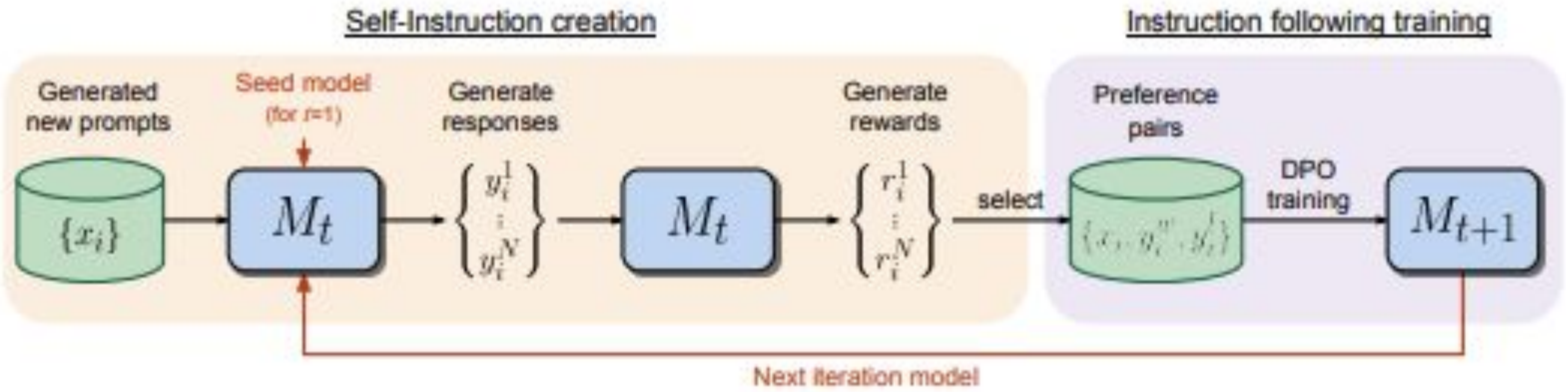
🏺: Suchit Gupte

Self-Rewarding Language Models

Self-Rewarding Language Models

Weizhe Yuan[1,2] Richard Yuanzhe Pang[1,2] Kyunghyun Cho[2] Xian Li[1] Sainbayar Sukhbaatar[1] Jing Xu[1] Jason Weston[1,2]

[1] Meta [2] NYU



Ability to generate and evaluate new instruction following examples to add to its own training set.

Given a prompt that describes a user request, the ability to generate a high quality, helpful (and harmless) response

THE OHIO STATE UNIVERSITY

Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

Zixiang Chen [*1] **Yihe Deng** [*1] **Huizhuo Yuan** [*1] **Kaixuan Ji** [1] **Quanquan Gu** [1]

- A new fine-tuning method called Self-Play fIne-tuNing (**SPIN**).
- SPIN starts from a supervised fine-tuned model.
- At the heart of SPIN lies a self-play mechanism – LLM generates its own training data from its previous iterations, refining its policy by discerning the self-generated responses obtained from the human-annotated data.
- Unlike the original work, which necessitates both a weak supervisor and a strong model, SPIN operates effectively with a single LLM.

Weak to Strong Generalization

Visionary

Alex Felderean

Improving Weak-To-Strong Generalization

Focus: demo analysis techniques for future “superhuman”-level models.

Need to see if this principle will be scalable!

Paper saw with current weak-to-strong generalization that generalization:

- disagrees with weak supervision when weak’s wrong.
- should not need too much modification to get desired.
- should be consistent between many prompts

Can we look at furthering the current generalization to better specify and test these requirements?

Identify New Unsupervised Properties

- Look into existing methods in ML literature to improve gains in generalization.

Better weak-to-strong generalization

=

Stronger ability to refine desired generalization for future stronger models.

- Refine scalable oversight methods to improve quality of weak supervisor.

Parallels to Semi-Supervised Learning

- Can employ when a small subset of labeled data (like in supervised learning) is available from a larger amount of unlabeled data (unsupervised learning).

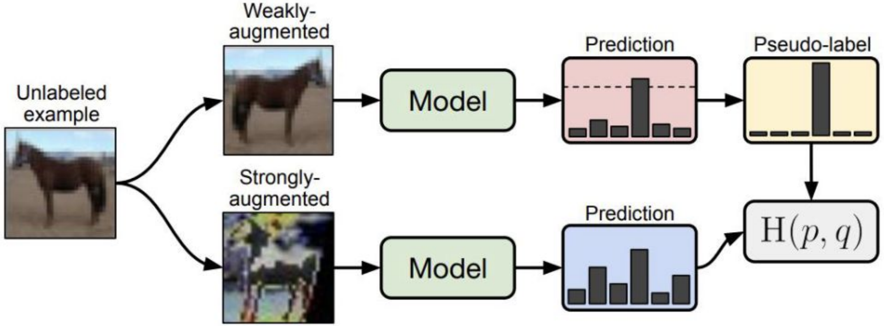
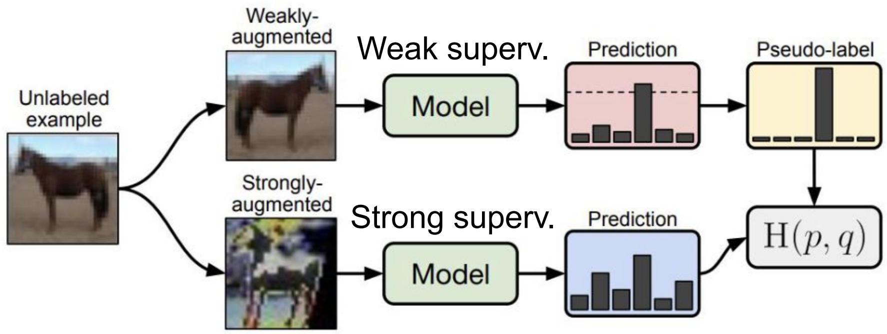


Figure 1: Diagram of FixMatch. A weakly-augmented image (top) is fed into the model to obtain predictions (red box). When the model assigns a probability to any class which is above a threshold (dotted line), the prediction is converted to a one-hot pseudo-label. Then, we compute the model's prediction for a strong augmentation of the same image (bottom). The model is trained to make its prediction on the strongly-augmented version match the pseudo-label via a cross-entropy loss.

Reference: https://doi.org/10.48550/arXiv.2001.07685

Proposed Vision of Incorporating Semi-Supervision

For some weak model-provided input, feed this to models with respective augmentation to string and label outputs on some labeling vector (ex. Sentiment / positivity, subject / topic, ethical, etc.), perform loss evaluation. Repeat for many different vectors and use loss functions to determine weak-strong agreement.



Reference: https://doi.org/10.48550/arXiv.2001.07685

Thank you