

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

Author: Burns et al.

created by paper2slides

Uploaded to arXiv: 2023-12-14

<https://arxiv.org/abs/2312.09390>

Executive Summary

- Study of aligning superhuman models using weak supervision through strong pretrained models.
- Investigation of naive finetuning with weak labels and identification of weak-to-strong generalization.
- Proposal of methods (e.g., confidence loss, bootstrapping) to enhance weak-to-strong generalization.
- Evaluation across NLP, chess, and ChatGPT reward modeling tasks to demonstrate traction.

Introduction

- Human evaluators struggle to supervise behavior of future superhuman models.
- Supervision needed to evaluate tasks like instruction adherence and safety.
- Proposal: Use weak models to supervise stronger models, analogous to humans supervising superhuman models.
- Core problem: Can weak supervision elicit full potential from stronger models?

Weak-to-Strong Learning

- Train strong models with labels generated by weak models.
- Measure performance across weak, strong student, and strong ceiling models.
- Assess robust methods to boost weak-to-strong generalization.
- Three tasks: NLP, chess, and ChatGPT reward modeling.

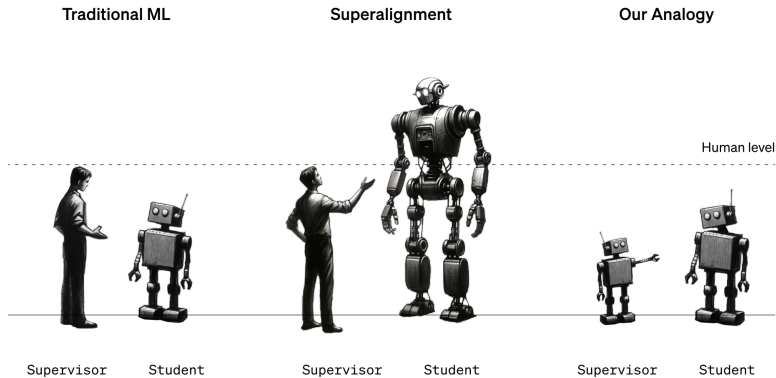


Figure: Experimental Setup: Using weak models to supervise strong models.

Task Details

- **NLP Tasks:** 22 popular binary classification datasets; e.g., sentiment analysis, commonsense reasoning.
- **Chess Puzzles:** Lichess dataset with chess positions; predict the best move.
- **ChatGPT Reward Modeling:** Predict human preferences between model responses.

Baseline: Naive Finetuning on Weak Labels

- Pretrained models from GPT-4 family.
- Naive finetuning results in weak-to-strong generalization.
- Demonstrated significant PGR even with large compute gap.

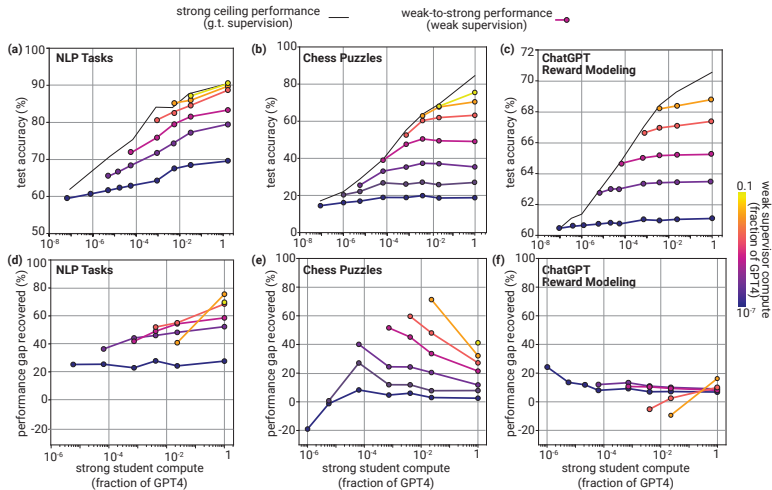


Figure: Promising generalization in NLP tasks but poor for ChatGPT RM.

Bootstrapping with Intermediate Model Sizes

- Sequential training approach.
- Intermediate models to bridge gaps between weak and strong models.
- Improved generalization in chess puzzles, little effect in NLP and RM.

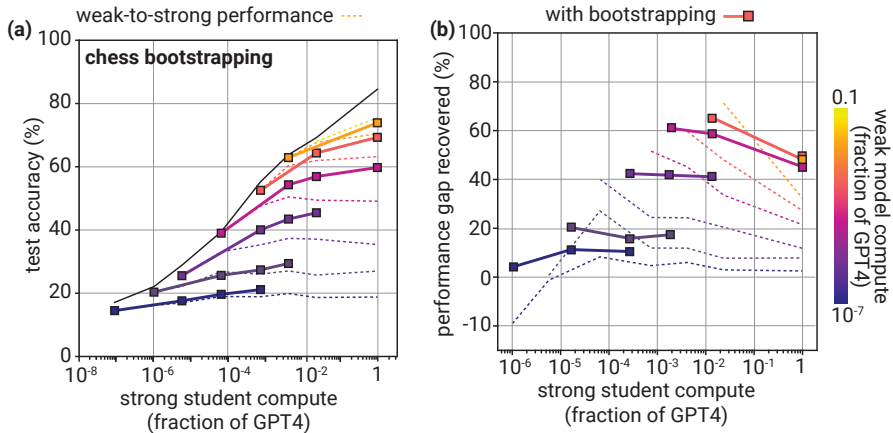


Figure: Bootstrapping improved generalization in chess puzzles.

Auxiliary Confidence Loss

- Encourages strong models to predict confidently, even when disagreeing with weak labels.
- Reduces imitation of weak supervisor errors.
- Drastically improves weak-to-strong generalization in NLP tasks.

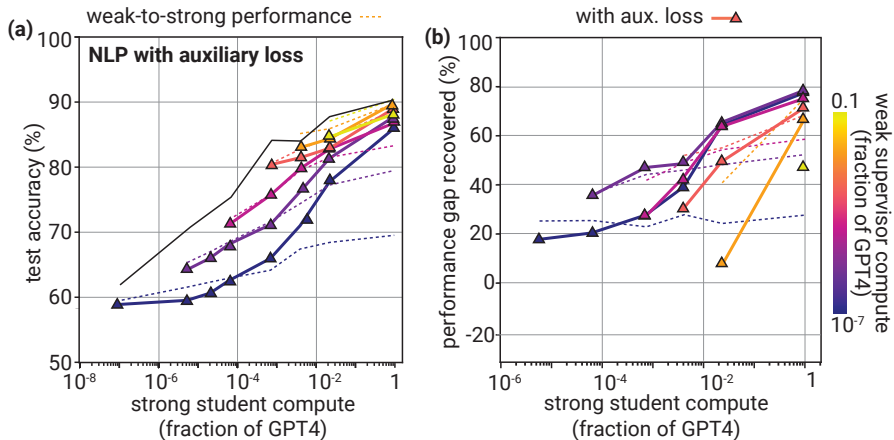


Figure: Substantially improved generalization on NLP datasets using confidence loss.

Generative Finetuning in ChatGPT RM

- Additional unsupervised finetuning step with language modeling objective.
- Increases ground truth concept saliency.
- Improves weak-to-strong PGR by 10-20% even with ceiling adjustments.

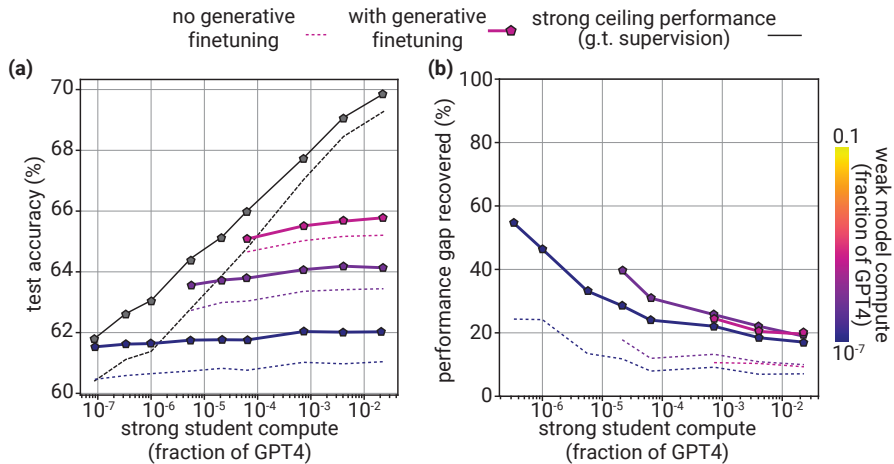


Figure: Generative finetuning improves performance and PGR in ChatGPT RM.

Understanding Generalization: Imitation

- Naive finetuning can overfit to weak labels.
- Overfitting observed even within one epoch.
- Imitation saliency affects generalization reliability.

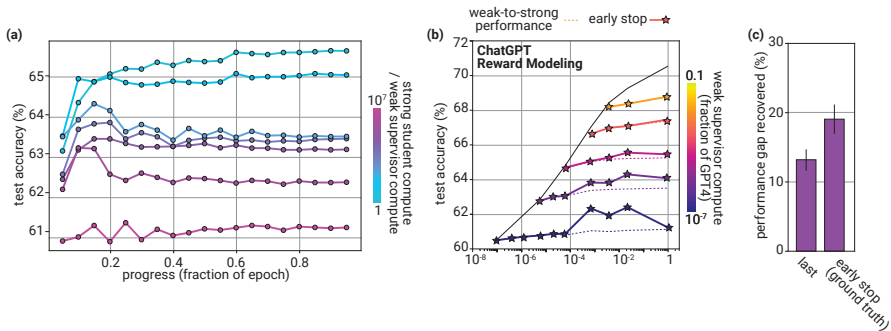


Figure: Overfitting to weak labels observed, indicating imitation.

Conclusion

- Weak-to-strong learning demonstrates promising results in multiple domains.
- Bootstrapping and confidence loss improve generalization.
- Future directions: Fix remaining disanalogies and improve scalable oversight methods.
- Necessity of strong understanding and validation of methods for high-stakes settings is critical.