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.

1 INTRODUCTION

We mainly steer or *align* today's models with reinforcement learning from human feedback (RLHF): we reinforce behaviors that human evaluators rate highly and penalize behaviors that evaluators rate poorly (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022a). This procedure is very effective when human evaluators can tell if model behavior is good or bad and is a core part of training modern language model assistants such as ChatGPT.

However, superhuman models will be capable of complex and creative behaviors that humans cannot fully understand. For example, if a superhuman assistant model generates a million lines of extremely complicated code, humans will not be able to provide reliable supervision for key alignment-relevant tasks, including: whether the code follows the user's intentions, whether the assistant model answers questions about the code honestly, whether the code is safe or dangerous to execute, and so on. As a result, if we finetune a superhuman model with human supervision on a reward modeling (RM) or safety classification task, it is unclear how that model will generalize to complicated behaviors that humans could not reliably supervise themselves.

This leads to a fundamental technical challenge of aligning superhuman models (superalignment): how can weak supervisors control models much smarter than them? Despite the importance of

*Primary authors. This was a joint project of the Superalignment Generalization team. Correspondence to generalization@openai.com. Code is available at github.com/openai/weak-to-strong.

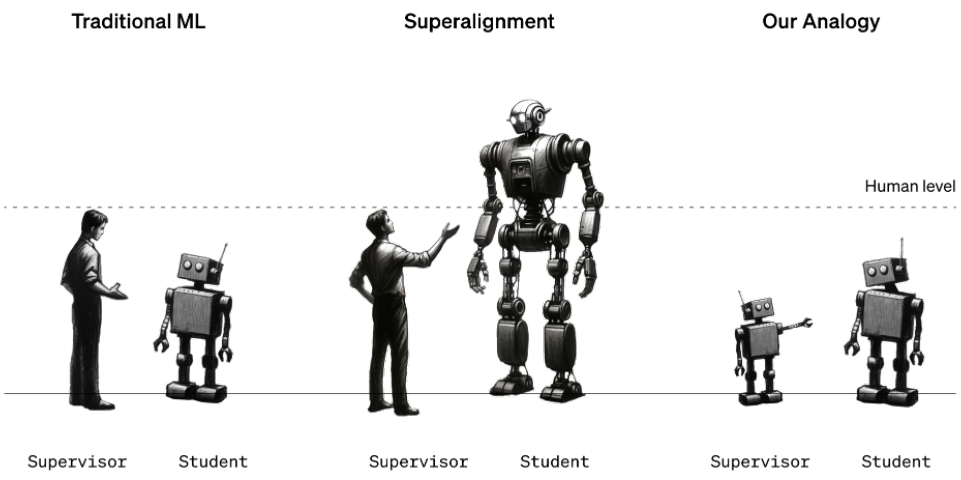


Figure 1: **An illustration of our methodology.** Traditional ML focuses on the setting where humans supervise models that are weaker than humans. For the ultimate superalignment problem, humans will have to supervise models much smarter than them. We study an analogous problem today: using weak models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

this problem, it is difficult to empirically study today. Most prior work on alignment has either confronted this core challenge head-on—but been restricted to primarily theoretical frameworks and toy problems (Irving et al., 2018; Christiano et al., 2018; Leike et al., 2018; Demski & Garrabrant, 2019; Hubinger et al., 2019), or empirically studied humans supervising today's models—without addressing the core challenges that may arise with superhuman models (Christiano et al., 2017; Wu et al., 2021; Ouyang et al., 2022; Bowman et al., 2022; Saunders et al., 2022). In contrast, we would ideally like to have a setup that captures core challenges of aligning future superhuman models while *also* being able to make iterative empirical progress today.

We propose a simple setup for studying the problem of humans supervising superhuman models by considering an analogy: can we use *weak models* to supervise *strong models*? We can empirically test this by finetuning large (strong) pretrained models on labels generated by small (weak) models and observing how they generalize. Just like the problem of humans supervising superhuman models, our setup is an instance of what we call the *weak-to-strong learning* problem.

Why should weak-to-strong learning be possible? On the one hand, the strong model could simply learn to imitate the weak supervisor, including its errors, since that is what we would naively train it to do. On the other hand, strong pretrained models should already have good representations of the alignment-relevant tasks we care about. 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*. This gives us hope that the strong model can generalize beyond the weak supervision, solving even hard problems for which the weak supervisor can only give incomplete or flawed training labels. We call this phenomenon *weak-to-strong generalization*.

We study our weak-to-strong learning setup (Section 3) by finetuning base (i.e. pretrained-only) language models from the GPT-4 family (OpenAI, 2023),[1] spanning 7 orders of magnitude (OOMs) of pretraining compute, across three settings: a large set of popular natural language processing (NLP) benchmarks, chess puzzles, and our internal ChatGPT reward modeling dataset. Our main findings include:

[1]These models share the same general architecture and pretraining dataset as GPT-4. However, this model series does not include the models known as GPT-2, GPT-3, and GPT-3.5.

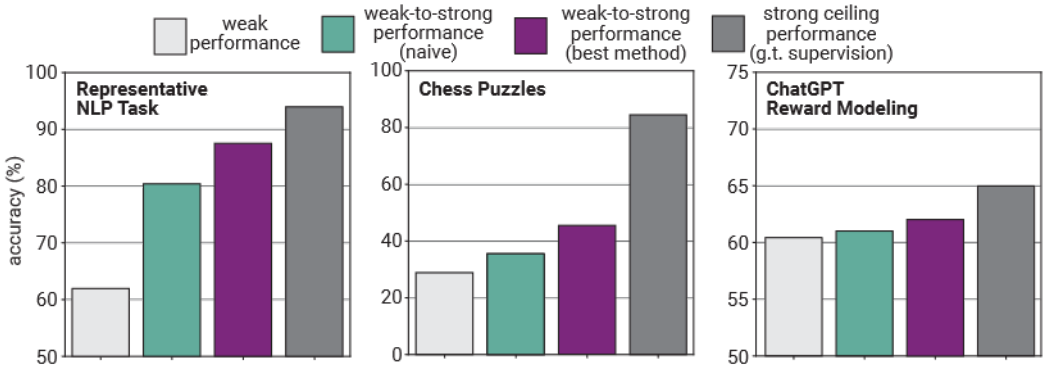


Figure 2: **Strong models trained with weak supervision generalize beyond their supervisor, and improving weak-to-strong generalization is tractable.** We show test accuracy on a representative NLP task (left), chess puzzles (middle) and the ChatGPT reward modeling task (right). We show the weak supervisor trained on ground truth labels (light grey) and the strong student trained with weak supervision naively (green), with the best method in each setting (purple), or with ground truth supervision (dark grey). For NLP and chess we supervise GPT-4 using GPT-2-level supervision, while for reward modeling we supervise a 3.5-level model using GPT-2-level supervision. The best method is the auxiliary confidence loss for the NLP task (Section 4.3.2), bootstrapping for Chess puzzles (Section 4.3.1), and unsupervised generative finetuning for reward modeling (Section 5.2.2; generative-finetuning is also used for the strong ceiling performance).

1. **Strong pretrained models naturally generalize beyond their weak supervisors.** If we naively finetune strong models with labels generated by weak models, they consistently outperform their weak supervisors (Section 4.2). For example, on NLP tasks, if we finetune GPT-4 with labels from a GPT-2-level model, we typically recover about half of the performance gap between the two models.

2. **Naively finetuning on weak supervison is not enough.** Despite positive weak-to-strong generalization, there still remains a substantial gap between strong models finetuned with weak supervision and strong models finetuned with ground truth supervision. Weak-to-strong generalization is particularly poor for ChatGPT reward modeling. Collectively, our results provide empirical evidence that naive RLHF will likely scale poorly to superhuman models without additional work.

3. **Improving weak-to-strong generalization is tractable.** We find that we can improve performance by encouraging strong models to have confident predictions with an auxiliary loss, bootstrapping supervision with intermediate models, and improving model representations with unsupervised finetuning. For example, when supervising GPT-4 with a GPT-2-level model on NLP tasks using the auxiliary confidence loss, we typically recover nearly 80% of the performance gap between the weak and strong models.

Our work has important limitations. None of our methods work consistently in all settings, and especially in the RM setting we are still far from recovering the full performance gap between weak and strong models. Thus our methods serve more as proofs-of-concept that weak-to-strong generalization is tractable, rather than practical solutions we recommend deploying today. Furthermore, there are still important disanalogies between our empirical setup and aligning superhuman models that we did not address (Section 6); continuously refining our basic setup will be important for ensuring that research today continues to make real progress toward aligning the superhuman models we develop in the future.

Despite the limitations of our work, we find our results to be highly encouraging. We show that substantial weak-to-strong generalization is not only possible, but actually a widespread phenomenon. We also show that with very simple methods, we can drastically improve the ability of weak supervisors to elicit knowledge from strong models. With much more progress in this direction, we could get to the point where we can use weak supervisors to reliably elicit knowledge from much stronger

models, at least for some key tasks that we care about. This may allow us to develop superhuman reward models or safety classifiers, which we could in turn use to align superhuman models.

Aligning superhuman models is essential for making them safe; there is increasing recognition that failing to align such powerful models has the potential to be catastrophic, making this one of the most important unsolved technical problems in the world (CAIS, 2022). We think it is now more tractable than ever to make rapid iterative empirical progress toward solving this problem.

2 Related Work

We study how we can leverage the generalization properties of deep neural networks to solve weak-to-strong learning. Our problem setting and methods are closely connected to many existing research areas.

Weakly-supervised learning. Weak-to-strong learning is a special type of weakly supervised learning—a setting in which models are trained using unreliable labels (Bach et al., 2017; Ratner et al., 2017; Guo et al., 2018). There is also a rich literature on the related problem of learning from noisy labels (Song et al., 2022). Common methods include bootstrapping (Reed et al., 2014; Han et al., 2018; Li et al., 2020), noise-robust losses (Zhang & Sabuncu, 2018; Hendrycks et al., 2018; Ma et al., 2020), and noise modeling (Yi & Wu, 2019). Unlike most work on label noise, the errors in our weak supervision are much harder to address than uniform label noise, instead having "instance-dependent" errors (Frénay & Verleysen, 2013). Semi-supervised learning, in which labels are only available for a subset of the data, is also closely related (Kingma et al., 2014; Laine & Aila, 2016; Berthelot et al., 2019). We could also study our problem in a semi-supervised setting by having an "easy" subset of examples that weak supervisors provide reliable labels for and a subset of unlabeled "hard" examples that the weak supervisor can't reliably label, a problem which we call "easy-to-hard generalization" (see Appendix C).

Student-teacher training. The framework of first training a teacher and then training a student on teacher's pseudo-labels is widely used in semi-supervised learning (Laine & Aila, 2016; Tarvainen & Valpola, 2017; Xie et al., 2020), domain adaptation (French et al., 2017; Shu et al., 2018), and knowledge distillation (Hinton et al., 2015; Gou et al., 2021; Stanton et al., 2021; Beyer et al., 2022). In contrast to most prior work, we focus on the setting where the student is much more capable than the teacher.

Furlanello et al. (2018) and Xie et al. (2020) also consider cases where the student is at least as capable as the teacher. However in their settings the student is randomly initialized and has access to ground truth labels. Moreover, compared to most past work we are focused on qualitatively *very* weak supervision. For example, we are interested in huge leaps in generalization, similar to going from "3rd grade-level" supervisors to "12th grade-level" student models. Despite these differences with past work, we expect many methods from semi-supervised learning and domain adaptation to translate to our setting. For example, we found that a type of confidence auxiliary loss similar to past work (Grandvalet & Bengio, 2004) improves weak-to-strong generalization in Section 4.3.

Robustness of pretraining and finetuning. Many papers have shown that pretraining on massive, diverse data leads to more robust representations that generalize better out-of-distribution (Hendrycks et al., 2019; 2020b; Radford et al., 2021; Liu et al., 2022). Finetuning typically improves in-distribution generalization, but often performs poorly out-of-distribution, sometimes even degrading performance relative to zero-shot prompting (Kumar et al., 2022; Wortsman et al., 2022b; Awadalla et al., 2022). Recent approaches to mitigating this problem include weight ensembling (Wortsman et al., 2022b;a), finetuning only a subset of layers (Kirichenko et al., 2023; Lee et al., 2022a), or mitigating the distortion effects that finetuning has on pretrained features (Kumar et al., 2022). We did not find strong results in preliminary explorations of approaches similar to these (Appendix B), but we expect that with more thorough explorations one may be able to attain much stronger results with these or other ideas from the robust finetuning literature.

Debiasing. In weak-to-strong generalization, the weak labels contain a specific form of bias, which results from the weak models' lack of capability. There is a substantial literature on learning from biased training data (Bellamy et al., 2018). However, most work focuses on *known* biases, for example where we know that the models perform worse on minority groups. For known biases, common methods include Group Distributionally Robust Optimization (Sagawa et al., 2019), adver-

sarial training (Zhang et al., 2018), and model editing (Santurkar et al., 2021; Meng et al., 2022). In contrast, our setting can be viewed as a particularly difficult debiasing problem where the bias is unknown. Some methods that automatically discover and mitigate biases include clustering (Sohoni et al., 2020), loss variance reduction (Khani et al., 2019), and auditing and re-training on high-loss group (Kim et al., 2019; Liu et al., 2021).

Imitation and preference learning. The goal of alignment is to steer already-capable models to do what we want them to do. For example, the base GPT-4 model is good at generating text following its pretraining distribution, but does not readily follow instructions. To align pretrained language models today, we finetune them using imitation learning on human demonstrations (Bain & Sammut, 1995; Atkeson & Schaal, 1997) or by using methods such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Glaese et al., 2022; Bai et al., 2022a). Constitutional AI (Bai et al., 2022b; Lee et al., 2023) leverages AI feedback to align language models, but still uses an initial RLHF phase. However, both imitation learning and preference learning assume high-quality human supervision, making it unclear if they will work for superhuman models.

Scalable oversight. Scalable oversight techniques aim to improve the ability of humans to supervise models. For example, humans may ask models to critique the outputs of other models (Irving et al., 2018; Saunders et al., 2022) or use models to help decompose a problem into simpler subproblems (Leike et al., 2018; Christiano et al., 2018; Lightman et al., 2023). Scalable oversight methods typically take advantage of special problem structure, like decomposability or the fact that evaluation is easier than generation. In contrast to improving human supervision, we focus on generalizing beyond human supervision such that models perform well even in settings we cannot reliably supervise. That said, our weak-to-strong learning setup can be used to compare scalable oversight methods, generalization-based methods, and more. Our setup also resembles a proposal for measuring progress on scalable oversight known as "sandwiching", which uses weak and strong humans (Cotra, 2021; Bowman, 2022).

Knowledge elicitation and honesty. Christiano et al. (2022) introduced a theoretical problem called Eliciting Latent Knowledge (ELK), in which the goal is to elicit latent knowledge from a superhuman machine learning model even under worst case assumptions. For example, a special case of ELK is honesty (Evans et al., 2021), where the goal is for the models to report their true beliefs[2]. Wentworth (2020) hypothesizes a tendency for neural networks to develop "natural abstractions" that are easier to elicit. Recent empirical work on ELK includes a benchmark for measurement tampering (Roger et al., 2023), methods for discovering latent knowledge (Burns et al., 2023), and studies of honesty (Li et al., 2023; Pacchiardi et al., 2023). Our setting can be viewed as a general methodology for empirically studying problems like ELK and honesty across a wide range of tasks.

3 Methodology

A core challenge of superalignment is that humans will need to supervise models much smarter than us. This is a special case of what we call the *weak-to-strong learning problem*: how can a weak supervisor oversee a model much smarter than it? In this paper, we study a simple analogy, in which we replace the weak human supervisor with a weak model supervisor.

For a given task of interest, consisting of a dataset and a performance metric, we:

1. **Create the weak supervisor.** Throughout most of this work, we create weak supervisors by finetuning small pretrained models on ground truth labels.[3] 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*.

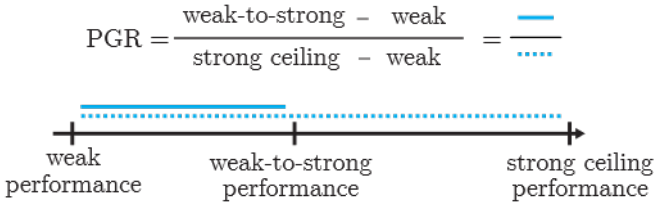
[2]Like Evans et al. (2021), we define *honesty* to mean a model reporting what it *believes* to be true, in contrast to truthfulness which asks whether what a model reports *is* true.

[3]In Appendix D and Appendix E we study other synthetic weak supervisors. Future work could test many more sources of weak supervision, such as by having 3rd grader humans provide labels.

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 performance 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.

For more details on how we train each model, see Appendix A.

Typically, weak-to-strong performance will be between weak performance and strong ceiling performance. We define the **performance gap recovered (PGR)** as a function of the above three performances (weak, weak-to-strong, and strong ceiling) as shown in the illustration below.



PGR measures the fraction of the performance gap (the difference in performance between the weak and strong ceiling models) that we can recover with weak supervision. If we achieve perfect weak-to-strong generalization, PGR is 1. If the weak-to-strong model does no better than the weak supervisor, then PGR is 0.

Advantages. Our setup has a number of advantages, including:

1. It can be studied with any pair of weak and strong models, making it easy to study scaling laws and not requiring access to expensive state-of-the-art models. Moreover, it does not require working with humans, so feedback loops are fast.
2. It can be studied for any task of interest, making it easy to empirically test across a wide range of settings.
3. Success will be practically useful even before we develop superhuman models: for example, if we find ways to align GPT-4 with only weak human supervision or with only GPT-3-level supervision, that would make it more convenient to align models today.

Limitations. Our setup still has important disanalogies to the ultimate problem of aligning superhuman models. We view our setup as removing one of the main disanalogies in prior work, not as providing a final, perfectly analogous setup. Two remaining disanalogies include:

1. **Imitation saliency.** Future superhuman models will likely have salient representations of human behaviors, but our strong models may not have learned features relevant for imitating weak model predictions; simply imitating the weak supervisor may thus be an easier failure mode to avoid in our setting than it will be in the future. More generally, the types of errors weak models make today may be different from the types of errors humans will make when attempting to supervise superhuman models.
2. **Pretraining leakage.** Our pretraining data implicitly contains supervision from humans. It may thus be artificially easy to elicit strong models' capabilities in our setting, since they were directly pretrained to observe strong (human-level) performance. Superhuman-level performance may not be directly observed in the same way—superhuman knowledge might be more latent, e.g. because it was learned from self-supervised learning—and thus might be harder to elicit from superhuman models in the future.

[4]For tasks solved by superhuman models that humans cannot evaluate, we will not have access to ground truth labels. However, we allow access to ground truth labels in our experimental setting today for scientific and evaluation purposes. Note that we evaluated weak-to-strong performance against ground truth many times while iterating on methods; however, we held out our largest model (GPT-4) and about half of NLP tasks throughout the project.

More generally, we do not yet know how superhuman models will be built, but they could develop new inductive biases that are qualitatively different from today's models. We view iterating on our methodology to produce even more analogous setups as a key priority for future work, as we discuss in more detail in Section 6.

4 Main Results

In this section, we report our main empirical results, including baselines and promising methods.

4.1 Tasks

Popular natural language processing benchmarks. We consider 22 popular NLP classification datasets covering ethics, commonsense reasoning, natural language inference, sentiment analysis, and other domains. We convert all datasets to binary classification tasks and approximately balance the classes. We produce soft labels from the weak model. See a full list of the datasets and their sources in Table 1.

Chess puzzles. We use the dataset originally introduced in Schwarzschild et al. (2021b), which contains chess puzzles from the `lichess.org` website (Lichess Team, 2023). Each puzzle consists of a chess position, and a sequence of optimal moves to play to solve the puzzle. For our evaluation, we predict the first move played, which is the best move in the given chess position. We illustrate the data format in Appendix Figure 14. For weak labels, we sample from the weak model with temperature 0. Note that unlike the other binary classification tasks we study in this paper, this is a generative task.

ChatGPT reward modeling. The standard approach to aligning models today is reinforcement learning from human feedback (RLHF). A critical step of RLHF is to train a reward model (RM) to predict human preferences between model responses. Specifically, a reward model is trained on a dataset consisting of dialogs between a human and an assistant model. For each query, the humans compare multiple possible responses (completions) from the assistant, providing human preference data. Then, a reward model is trained to predict the results of pairwise comparisons between completions. Finally, the assistant model is trained by optimizing against the reward model with reinforcement learning (RL). In our work, we do not study the RL step, and instead assume the goal is to maximize reward model accuracy. For more details on reward models, see e.g. Ouyang et al. (2022). We use a proprietary dataset used to train ChatGPT reward models.

For more details about our tasks and setup, see Appendix A.

4.2 Naively finetuning on weak labels

In each of these 3 settings (NLP tasks, chess puzzles, and reward modeling) we evaluate how well strong students generalize when naively finetuned on labels generated by weak supervisors. We study pretrained language models from the GPT-4 family (OpenAI, 2023), which allow us to study student-supervisor compute disparities of many orders of magnitude. We find that PGRs are almost universally positive—in virtually all settings that we studied, and across almost all student and supervisor sizes, students outperform their supervisors (Figure 3).

On the popular NLP benchmarks, we find especially promising weak-to-strong generalization: strong models trained with weak supervision can often generalize to a substantially higher performance than the weak model itself. Even with very weak supervisors and strong models with many orders of magnitude more compute, we recover more than 20% of the performance gap. The PGR increases both with weak supervisor size and with strong student size; for the largest students, the PGR is often above 50%.

We see more mixed results in the chess puzzle setting. In particular, when using the smallest weak models, the PGR is close to zero and the test accuracy curves appear flat. However, as the size of the weak supervisor increases, the PGR increases substantially; for small supervisor-student gaps, PGR can be above 40%. Unlike in the NLP setting, where PGR improves with the strong student size, PGR *decreases* with the strong student size for a given weak supervisor on chess puzzles. The cor-

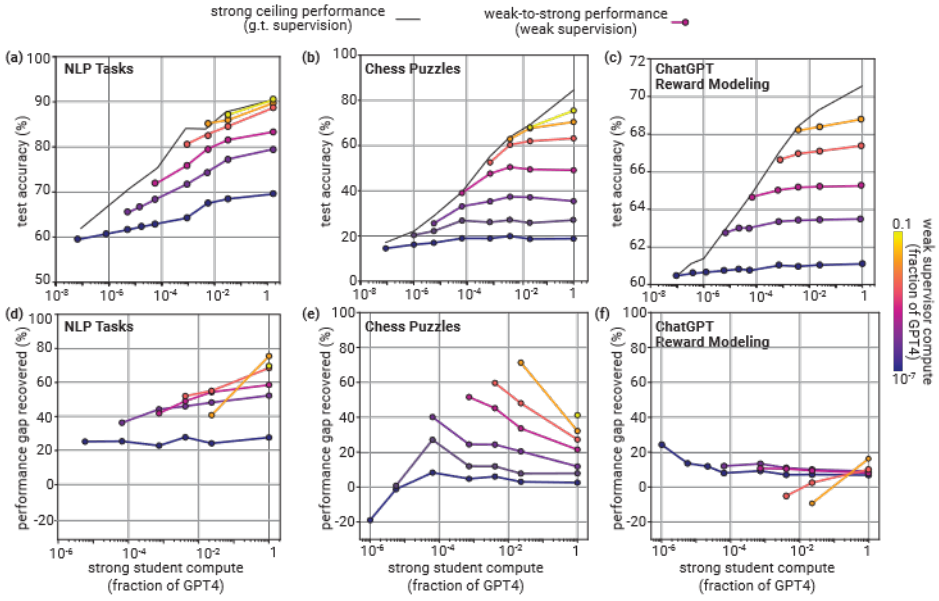


Figure 3: **Promising weak-to-strong generalization with naive finetuning on NLP tasks and chess, but poor generalization on the ChatGPT reward modeling task. (a,b,c)** Test accuracy as a function of strong student size on **(a)** NLP tasks, **(b)** chess puzzles, and **(c)** the ChatGPT reward modeling task. Accuracy of strong students trained with ground truth in black, accuracy of strong students trained with weak supervision shown with colored lines (hue indicates size of weak supervisor). **(d,e,f)** Same as panels **a,b,c** but for performance gap recovered (see Section 3 for details). For NLP settings, we compute the median across tasks (see Figure 12 for full details). We find decent weak-to-strong generalization and even positive PGR scaling on NLP tasks, decent generalization for small supervisor-student gaps but negative PGR scaling on chess puzzles, and both poor generalization and scaling for ChatGPT reward modeling.

responding test accuracy curves appear concave, potentially exhibiting inverse scaling (McKenzie et al., 2023) in strong student size.

Finally, we find that weak-to-strong generalization is poor by default in the ChatGPT reward model setting. We are usually only able to recover roughly 10% of the performance gap between the weak supervisor and the strong student. Even for relatively small gaps in compute between the weak and strong models, PGR almost never exceeds 20%.

In general, across all our settings, we observe weak-to-strong generalization: strong students consistently outperform their weak supervisors. It is not obvious why this should happen at all—especially from naive finetuning alone—and it gives us hope that weak-to-strong learning is a tractable problem. At the same time, our results suggest that naively using weak, human-level supervision will be insufficient to align strong, superhuman models; we will need qualitatively new techniques to solve superalignment.

4.3 Improving Weak-to-Strong Generalization is Tractable

We now show that we can use simple methods to substantially improve weak-to-strong generalization. While none of the methods we test works universally, these methods are proofs-of-concept that across many different tasks we can substantially improve generalization.

4.3.1 Bootstrapping with intermediate model sizes

Bootstrapping is a long-standing idea in alignment: instead of directly aligning very superhuman models, we could first align an only slightly superhuman model, use that to align an even smarter model, and so on (Christiano, 2019; 2018; Leike & Sutskever, 2023; Worley, 2021). Our setting allows us to empirically test this idea.

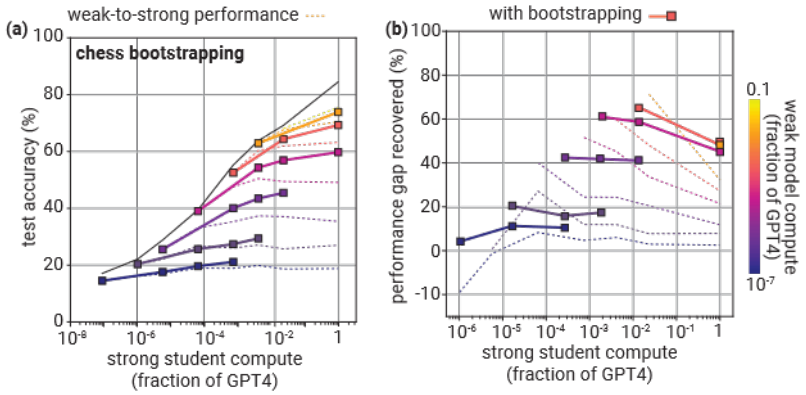


Figure 4: **Bootstrapping improves weak-to-strong generalization on chess puzzles.** **(a)** Test accuracy as a function of strong student size. Accuracy of students trained with ground truth in black, accuracy of students naively trained with weak supervision shown with dotted lines (hue indicates size of weak supervisor). Accuracies of students trained via bootstrapping shown with colored squares (including both the final weak-to-strong performance and the performance of the intermediate models during bootstrapping). **(b)** Same as **a** with PGR. By taking multiple small steps instead of one big step we see substantially improved generalization, especially for larger student models.

Specifically, we can construct a sequence of model sizes $\mathcal{M}_1 \to \mathcal{M}_2 \to \ldots \to \mathcal{M}_n$ of increasing sizes. Then, we use the weak labels from $\mathcal{M}_1$ to finetune $\mathcal{M}_2$, use $\mathcal{M}_2$ to generate new weak labels that we can use to finetune the next model in the sequence, $\mathcal{M}_3$, and so on.

We evaluate bootstrapping in the chess puzzle setting. When we naively finetune on weak labels for chess (Section 4.2), we see high PGR when we cross small supervisor-student gaps, but low PGR for larger gaps. As a result, in this setting it may help to take multiple small steps—steps where PGR should be high—instead of one big step.

For each round of bootstrapping, we run three iterations of weak-to-strong learning, i.e. we bootstrap the weak supervision using two intermediate model sizes before finally finetuning the largest model in the sequence. We report the results (including all intermediate weak-to-strong models within each bootstrap) in Figure 4. Bootstrapping improves PGR compared to the baseline, especially for larger student models. With the naive method, transfer accuracy curves flatten as the weak-strong gap grows larger; with bootstrapping, the accuracy continues to monotonically improve.

While the results in the chess setting are promising, in preliminary experiments we observed only small improvements with bootstrapping on NLP tasks and no improvements in the RM setting. This makes sense intuitively: unlike in the chess setting where naive PGR decreased with larger supervisor-student gaps, naive PGR increased or was rougly constant for larger supervisor-student gaps in the NLP and reward modeling settings. Overall, these results suggest bootstrapping is a plausible avenue to investigate for improving weak-to-strong generalization and can be helpful in some settings, but that naive bootstrapping alone will not be enough to align models much smarter than their supervisors.

4.3.2 An auxiliary confidence loss can dramatically improve generalization on NLP tasks

In our baseline results (Section 4.2), we naively finetune the strong student on the labels provided by the weak supervisor. Because we are directly training the strong student to imitate the weak supervisor, it may also learn to imitate the errors of the supervisor (see Section 5.1 for more discussion). Intuitively, we want to avoid this failure mode and provide additional regularization towards what the strong pretrained model already internally knows: we want the student to learn the intent of the supervisor, but not to imitate its mistakes.

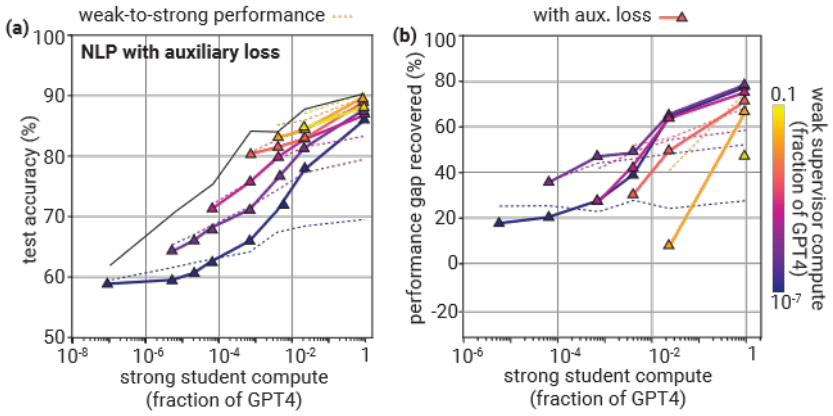


Figure 5: **Substantially improved generalization on NLP datasets with a simple auxiliary loss. (a)** Test accuracy as a function of strong student size. Accuracy of a student trained with ground truth in black, accuracy of students naively trained with weak supervision shown with dotted lines. Accuracies of students trained with auxiliary confidence loss shown with colored triangles. Median computed across 22 NLP tasks (hue indicates size of weak supervisor), see Figure 6 for individual datasets. **(b)** Same as **a** with PGR. The confidence loss can improve generalization drastically, especially for large supervisor-student gaps.

We operationalize this intuition by adding an auxiliary confidence loss term to the standard cross entropy objective. This method is closely related to conditional entropy minimization (Grandvalet & Bengio, 2004) which is a prominent technique in semi-supervised learning. Specifically, we add an additional loss term which reinforces the strong model's confidence in its own predictions—even when they disagree with the weak labels. We provide a detailed description of the method in Appendix A.4.

In Figure 5, we plot accuracy and PGR curves with this method on our NLP tasks. We find that while it performs slightly worse than the naive baseline for smaller strong students, it dramatically improves generalization for large gaps in compute between weak and strong models. With the smallest weak supervisor and largest strong student, the confidence loss increases median PGR from about 25% to nearly 80%.

In addition, we also plot generalization curves for a representative subset of NLP datasets in Figure 6, as well as the full panel of datasets in Figure 12. There are some settings in which the confidence loss does not help much or degrades performance, e.g. when the gap between the weak supervisor and strong student is small or when the dataset features inverse scaling even with ground truth supervision. But the confidence loss improves performance on most NLP datasets dramatically, and for many datasets we get almost perfect generalization, recovering nearly all the performance of the strong model, even when using the smallest weak supervisors.

Finally, we find evidence consistent with our motivating intuition for the confidence loss (allowing the strong student to confidently disagree with its weak supervisor): the auxiliary loss reduces the strong student's imitation of weak errors and mitigates weak label overfitting (see Section 5.1).

5 Understanding Weak-to-Strong Generalization

Strong methods will be essential for solving superalignment, but to trust those methods it is also important to understand *when* and *why* they work. A better understanding of weak-to-strong generalization could help us trust that generalization will continue working even in the future high-stakes settings we care most about, and could help us develop better methods along the way. In this section, we study two phenomena relevant to weak-to-strong generalization: imitation of supervisor mistakes and salience of the tasks to the strong student model.

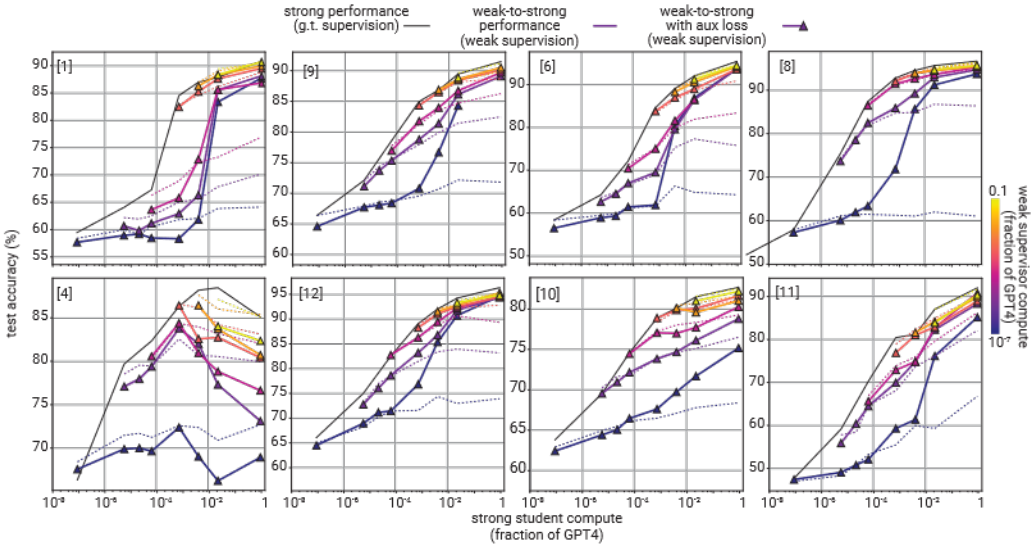


Figure 6: **Simple auxiliary loss improves generalization across most datasets.** Test accuracy as a function of strong student compute for a representative sample of NLP tasks. See Table 1 for dataset details and Appendix Figure 12 for results on all 22 NLP tasks. Auxiliary loss is shown with triangles, and the baseline with dotted lines. Weak supervisor model size shown in varying colors, with ground truth supervision shown in black.

5.1 Understanding imitation

When we train a strong model with weak supervision on some task, our hope is that the strong model will perform that desired task as well as possible, leveraging the latent capabilities it learned from pretraining to significantly outperform the weak supervisor. A salient way in which we could fail to achieve that desired generalization is if the strong model instead learns to imitate the weak supervisor—predicting how the weak supervisor would have classified each example. In particular, if the weak labels contain systematic errors that are easy to learn, the strong model could learn to imitate those errors. This is also a concern raised in theoretical work on superalignment, which has argued that the *human simulator* failure mode could be important: naive human supervision might result in superhuman models learning to imitate what a human would say, rather outputting its best predictions (Christiano et al., 2022).

5.1.1 Overfitting to Weak Supervision

The failure mode of imitating weak supervision is especially relevant to our naive baseline in Section 4.2, which directly trains the student to imitate the supervisor. In the case of infinite training data, naively fitting the weak labels should result in perfect imitation, and a PGR of zero. In practice, we train on finite data for a small number of epochs. Unlike typical ML settings, however, we could expect to observe overfitting even when training for less than a single epoch: the strong model might overfit to the weak supervisor labels and its errors, degrading ground truth test accuracy over training even without classic overfitting to any specific training examples.

Empirically, we see that the strong student indeed appears to overfit to the weak supervisor's errors. In Figure 7(a) we show ground truth test accuracy curves over the course of training for the ChatGPT RM task, and in Figure 7(b) and (c) we compare the best[5] and final ground truth test accuracies (median across all weak-strong model pairs). We find overfitting for large weak-strong gaps. For small weak-strong gaps, weak-to-strong performance typically monotonically increases over the course of training. For larger gaps, weak-to-strong performance often increases initially, but then starts dropping well before a single epoch has elapsed. Ground truth early stopping, which "cheats"

[5]Note that our best test accuracies may slightly overstate accuracy, due to noisy evaluations.

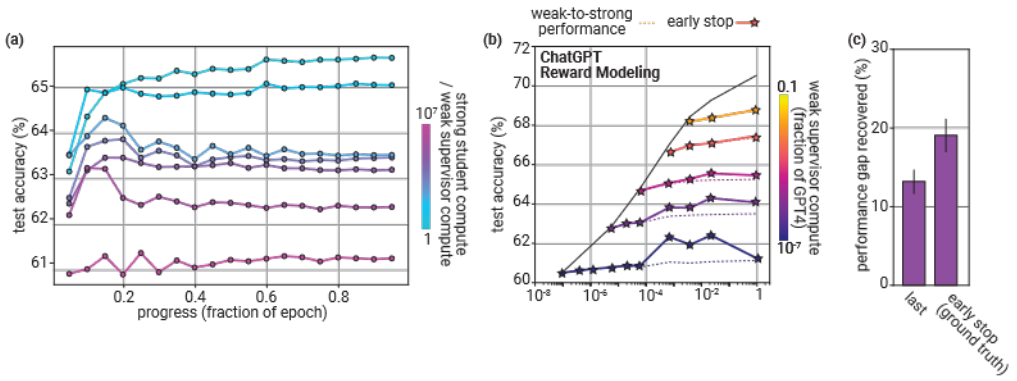


Figure 7: **Strong models overfit to the weak labels.** In all figures, we show data for the ChatGPT Reward Modeling task. **(a)** Weak-to-strong performance over the course of training. Hues indicate the student-supervisor gap. (**b**) Best weak-to-strong performance during training (stars) and weak-to-strong performance at the end of training (dashed). Weak performance in black. Hue indicates the size of the weak supervisor. (**c**) Median best and final performance gap recovered (PGR) aggregated across all supervisor-student pairs. We see overfitting to weak labels for large weak-strong gaps, even within one epoch. In these cases, the best test accuracy achieved over training can be substantially better than the test accuracy at the end of training. See Figure 13 for the corresponding analysis of a representative subset of NLP tasks.

by evaluating against ground truth and stopping at an optimal step with respect to ground truth test labels, typically gives a PGR improvement of around 5 percentage points.

We see the same phenomenon for NLP tasks in Figure 13. In the NLP setting, we find that "cheating" early stopping on ground truth gives a 15 percentage point boost in PGR over the model at the end of training, and a 10 percentage point boost in PGR compared to "non-cheating" early stopping with respect to weak labels.

Unfortunately, an early stopping criterion that uses ground truth labels does not constitute a valid method. Nevertheless, the results above suggest that imitating weak supervisor errors may be an important phenomenon in our setting.

Moreover, these results suggest that better early stopping or regularization strategies may be able to substantially improve weak-to-strong generalization, by reducing overfitting to the weak labels and their errors. Indeed, we see in Figure 13 that the auxiliary confidence loss introduced in Section 4.3.2 reduces overfitting to weak labels on NLP tasks substantially. For large weak-strong gaps, early stopping on ground truth (compared to early stopping on weak labels) gives a 15% PGR boost when using the naive method, but only a roughly 5% PGR boost when using the confidence loss.

5.1.2 Student-supervisor agreement

Another way to measure imitation is to directly measure the agreement between the student and the supervisor: the fraction of test inputs where the strong student makes the same prediction as the weak supervisor. Note that if agreement were 100%, then weak-to-strong accuracy would be equal to supervisor accuracy, and PGR would be 0.

In general, we notice that for our naive finetuning baseline, student-supervisor agreement is consistently high—often noticeably higher than weak supervisor accuracy. This indicates that the student is imitating some of the supervisor's errors. These phenomena hold across all tasks (NLP tasks, chess, and reward modeling) and all model sizes, for the naive method.

The confidence loss in Section 4.3.2 reduces student-supervisor agreements significantly (Figure 8), primarily by imitating supervisor mistakes less (Figure 8c). The loss encourages the strong student to make confident predictions, including when they contradict the weak supervisor. In a handful of the settings where it is most successful, the confidence loss reduces student-supervisor agreement

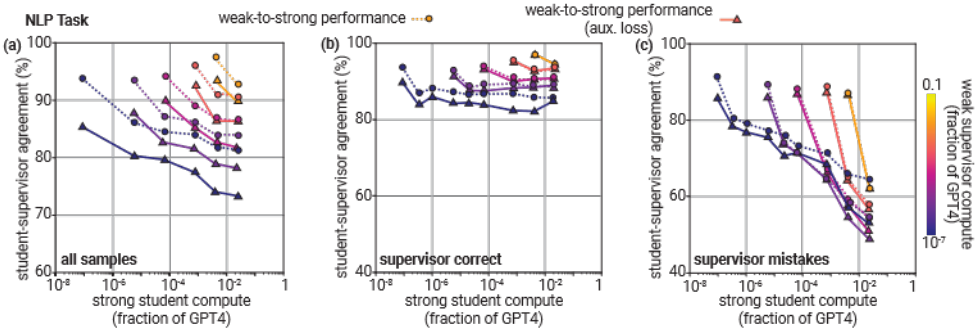


Figure 8: **Student-supervisor agreement *decreases* with larger student-supervisor gaps; the confidence loss reduces imitation of supervisor mistakes.** **(a)** Student-supervisor agreement as a function of strong student size on NLP tasks, **(b)** **a** but only on samples where the supervisor is correct, **(c)** **a** but only on samples where the supervisor is mistaken. Dotted lines indicate naive finetuning on weak labels, and triangles indicate results with the auxiliary confidence loss results (see Section 4.3). Hue of line indicates size of weak supervisor. For results on reward models, see Figure 16.

below strong student test accuracy (weak-to-strong performance)—i.e., the resulting model is fitting the ground truth concept *better* than it is fitting the weak labels it was trained with.

5.1.3 Inverse scaling for imitating the supervisor

Next, we study student-supervisor agreement as a function strong model size (see Figure 8 and Figure 16). Surprisingly, we find inverse scaling (McKenzie et al., 2023): larger student models consistently agree *less* with the errors of the supervisor than smaller student models, despite being trained to imitate the supervisor, not using early stopping, and having larger capacity than smaller student models.

This trend is especially strong if we evaluate agreement only on datapoints where the supervisor is wrong (Figure 8c), and the trend persists if looking at cross entropy loss instead of accuracy.

These results suggest that pretrained models may have a hard time fitting errors of other (smaller) pretrained models, at least in finetuning settings with relatively limited data. Stanton et al. (2021) and Furlanello et al. (2018) report a related observation in the context of knowledge distillation: it is surprisingly hard for models to fit the predictions of other models, even when they have sufficient capacity to do so.

One natural hypothesis is that the nature of (especially naive) weak-to-strong generalization depends heavily on the error structure of the weak supervisors and how easy those errors are to imitate. In Appendix E, we show initial experiments that test how different types of weak supervision errors impact what the strong student learns. Our results suggest that errors that are more difficult for the student to imitate result in stronger naive weak-to-strong generalization, but that even when they are easy to imitate, the confidence loss can help.

5.2 Saliency in the strong model representations

One intuition for when weak-to-strong generalization might be feasible is when the task or concept we want to elicit is internally "salient" to the strong model. In this section, we study several phenomena related to the saliency of the concepts we are trying to elicit from the student model.

5.2.1 Eliciting strong model knowledge with prompting

One possible reason for the high PGR we observe in Section 4 could be that eliciting what the strong model knows is easy. In particular, it is possible that strong pretrained models can solve many relevant tasks zero-shot with a simple prompt.

In Figure 9a, we consider 7 representative NLP tasks and compare finetuning, zero-shot prompting, and 5-shot prompting; for this initial experiment, we use ground truth labels rather than weak labels

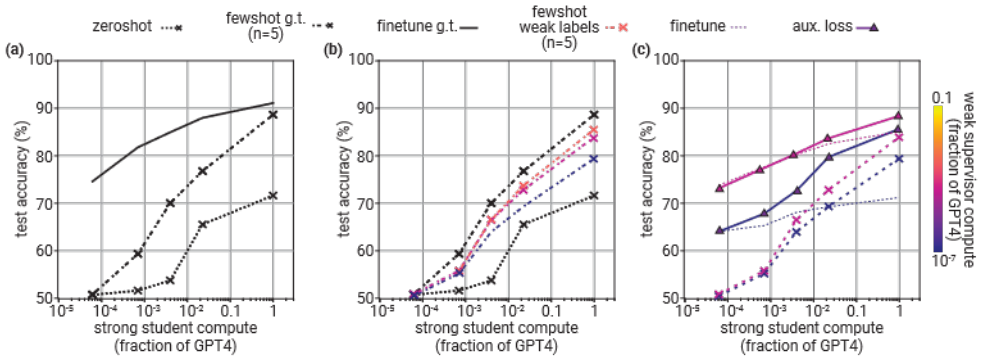


Figure 9: **Few-shot prompting becomes competitive with finetuning for large models; weak-to-strong learning is qualitatively similar in the prompting setting. (a)** Average zero-shot (single dashed), 5-shot (double dashed) and finetuning (solid) accuracy with ground truth labels as a function of strong student size. **(b)** Average 5-shot with weak labels (colored dashed) accuracy as a function of student model size. Hue of line indicates size of weak supervisor. Zero-shot and 5-shot same as in panel **a**. **(c)** Average weak-to-strong performance for 5-shot prompting (dashed with crosses), naive finetuning (dashed thin) and finetuning with the confidence loss (solid with triangle) as a function of student model compute. Results are averaged across 7 NLP tasks. Few-shot weak-to-strong performance becomes competitive with or outperforms finetuning for the largest strong students, though finetuning with the confidence loss does better.

for finetuning and 5-shot. For both the zero-shot and 5-shot baseline we use task-specific prompts summarized in Table 2. We find that zero-shot and 5-shot test accuracy is poor for most model sizes but, consistent with Brown et al. (2020), improves drastically for larger model sizes. In particular, for the largest models, 5-shot prompting becomes competitive with finetuning on many tasks, indicating that eliciting the task-relevant knowledge of these very large models is relatively straightforward.

We are also interested in weak-to-strong learning in the context of few-shot prompting. To study this setting, we construct a few-shot prompt where the labels are provided by the weak supervisor. We report the results in Figure 9b. Consistent with our findings in the finetuning setting, we get worse performance when we few-shot prompt with weak labels than we do few-shot prompting with ground truth labels. This suggests that weak-to-strong learning is a nontrivial problem in the prompting setting as well.

Similar to the finetuning setting, few-shot weak-to-strong performance improves for stronger supervisors. Compared to our weak-to-strong finetuning baseline (Figure 9c), weak-to-strong performance of few-shot prompting is poor for smaller student models, but becomes competitive or even outperforms finetuning for the largest strong students. However, weak-to-strong finetuning with the confidence loss still generally outperforms weak-to-strong few-shot prompting.

Overall, these results provide an important reference for our results on weak-to-strong generalization. They suggest that for the largest model sizes, the knowledge needed to solve many task can be elicited fairly easily with prompting. However, our current setup may be more disanalogous for prompting than for finetuning; many of our NLP tasks may have been implicitly observed during pretraining, which we conjecture benefits prompting more than finetuning. We discuss this potential disanalogy much more in Section 6.1.

5.2.2 Generative supervision improves RM weak-to-strong generalization

If salient representations of the desired task is useful for weak-to-strong generalization, then we may be able to improve generalization by increasing the salience of the task to the strong model. One way to increase the salience of a task without needing ground truth labels is to perform unsupervised finetuning with the language modeling objective on data relevant to that task (Dai & Le, 2015). For example, by finetuning a language model in an unsupervised way on online reviews, sentiment becomes saliently represented to models internally (Radford et al., 2017).

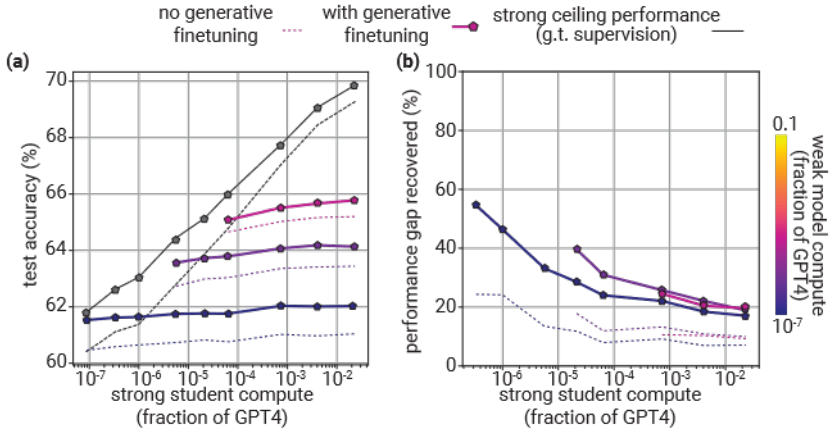


Figure 10: **Generative finetuning on reward modeling data improves weak-to-strong performance and PGR. (a)** Weak-to-strong performance on the reward modeling task, with (solid lines) and without (dashed lines) an extra step of generative finetuning for the strong student model. Solid black line shows a strong ceiling reward model that was also trained with the generative finetuning step; dashed black line show a weak supervisor reward model trained *without* the generative finetuning step. **(b)** PGR with and without generative finetuning. For generative finetuning PGR, we use the strong ceiling performance that also had this extra generative finetuning step. Even with this ceiling adjustment, PGR is higher with an extra generative finetuning step.

We test this idea in our reward modeling setting, where it is standard practice to initialize the model with a baseline finetuned on demonstrations of desired behaviors (Stiennon et al., 2020). In our case, we re-use the ChatGPT comparison data instead of introducing a new supervision dataset. Comparisons are comprised of a prefix (a single request or conversation between the user and assistant) and at least two candidate completions. We finetune the base models with a language modeling loss on all prefix-completion pairs, ignoring the human preferences between those completions.

Note that these pairs include completions ranked worst by human raters, so this procedure should not in principle leak any information about the ground truth preference labels that the weak-to-strong models should not have access to. On the other hand, since the completions can come from humans or stronger models, there may be some leakage similar in kind to the pretraining leakage that we discuss as a disanalogy in Section 6.1. Even in this setup, the reward modeling task is highly non-trivial, and we leave addressing this disanalogy (e.g. by collecting completions only from weaker models) for future work.

We found that the additional generative finetuning on the RM data leads to better weak-to-strong performance. Because this procedure also improves the performance of models trained on ground truth RM data, we compare our new weak-to-strong performance to strong "ceiling" models that were also first generatively finetuned in the same way. Even with this adjusted ceiling, we find that generative supervision improves PGR by approximately 10-20%. We report the results in Figure 10.

Furthermore, the improvement from generative finetuning stacks with the improvement from ground truth early-stopping (a "cheating" method to illustrate potential performance if we could optimally early stop, see Section 5.1.1). When we combine these two techniques, we can achieve PGR of approximately 30-40%, which would make the results on the RM task competitive with the weak-to-strong generalization we observe on NLP and chess puzzle tasks.

We can apply the idea of improving task saliency with generative finetuning on relevant data to all settings, and we believe this could be a promising direction for future work.

5.2.3 Finetuning on weak supervision to increase concept saliency

One possible measure of concept saliency is how linearly represented a task is. In particular, we can measure the performance of a linear probe (logistic regression classifier) trained from frozen activations of the model. If the optimal solution can be approximately recovered with a linear probe, that

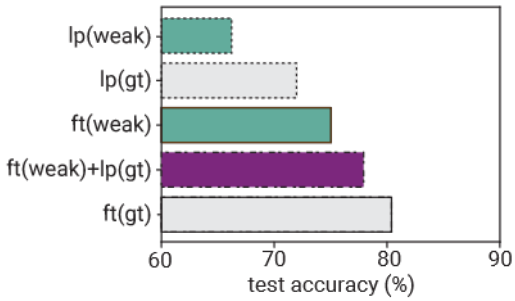


Figure 11: **Finetuning on weak supervisor labels makes the desired generalization more linearly represented.** We plot test accuracy for five different strategies, averaged across a subset of NLP tasks. **lp(weak)**: training a linear probe on the base model using weak labels, **lp(gt)**: training a linear probe on the base models using ground truth labels, **ft(weak)**: finetuning the model on weak labels, **ft(weak) + lp(gt)**: finetuning the model on weak labels *then* training a linear probe on ground truth labels, **ft(gt)**: finetuning the model on ground truth labels. Finetuning on the weak labels significantly increases the linearity of the ground truth concept.

could simplify our problem greatly; we could focus on linear probing methods instead of finetuning methods, which could greatly reduce the search space we need to consider to elicit the desired generalization. In our work, we focus only on how linearly represented a task is in the final activations, prior to the unembedding layer.

In Figure 11, we plot average test accuracy on a subset of our NLP datasets for several different combinations of (1) finetuning or linear probing, using (2) weak or ground truth labels. First, we show linear probes trained with ground truth labels (72% accuracy on average) perform worse than finetuning with ground truth labels (82% on average), indicating that the optimal solution to most tasks is *not* represented completely linearly in the strong model's final activations. For comparison, we also report the results for linear probing and finetuning using weak labels, which we verify are worse than using ground-truth labels.

However, we find that we can achieve substantially better performance by *first* finetuning the model on the *weak* labels, and *then* linear probing using the *ground truth* labels. In other words, when we finetune the strong model with weak labels, the representations become *more linear even with respect to ground truth labels*. In fact, finetuning on weak labels then linear probing on ground truth labels results in an accuracy of 78%, closing 60% of the gap between ground truth linear probing and finetuning. This also noticeably outperforms the naive weak-to-strong finetuning baseline.

This phenomenon is closely related to a recent finding reported by Kirichenko et al. (2023) in the spurious cues literature. They find that finetuning a model on biased supervision can result in models with very biased outputs, but surprisingly strong linear representations of the desired concepts. These results suggest an alternative approach to improving weak-to-strong generalization. We could first "linearize" the desired concept, e.g. by naively finetuning on weak labels. Then we could use simpler linear probe-based weak-to-strong methods to elicit the desired concept.

6 Discussion

In this paper, we proposed a simple analogy for studying a core challenge of aligning superhuman models and showed that it is feasible to make significant progress on this problem. However, our setup still has important disanalogies, which we now elaborate on. We then outline a number of promising avenues for future work.

6.1 Remaining disanalogies

Imitation saliency: superhuman models may easily imitate weak errors. Future models will likely be very good at predicting what humans will think and say, especially if they are trained on human data in a similar manner to current models. Consequently, if we naively train such a

superhuman model with human supervision, it might simply imitate the weak supervisor, outputting human-level capabilities rather than its latent superhuman capabilities (Christiano et al., 2022).

This problem is only partially captured by our setup. While our strong pretrained models do imitate weak supervisors to some extent, they are not explicitly pretrained to imitate weak models, and our results from Section 5.1.3 suggest that larger strong models may even have more difficulty doing this imitation. As such, "imitating the weak supervisor" may not be as much of a problem in our setup as it will be for the ultimate superalignment problem. This may inflate generalization performance today. We believe a more thorough investigation of this problem is an important area for future work.

Pretraining leakage: superhuman knowledge may be latent, not observable. Many of the tasks we consider in this work may have been observed in pretraining at least indirectly, for example through questions on online forums or through slight reframings of the task. For example, it is highly likely that simple science questions similar to those in the SciQ NLP task are present in our GPT-4 series pretraining dataset at least implicitly in some form. However future superhuman models may never directly observe superhuman alignment-relevant capabilities; these capabilities may be predominantly "latent", e.g. learned through self-supervised learning or reinforcement learning rather than through imitation learning. Intuitively, latent capabilities may be harder to elicit than capabilities that models could have observed in their pretraining data.

This disanalogy could cause our results to be overly optimistic. We conjecture that this disanalogy also increases prompting performance (Section 5.2.1) more than it increases finetuning performance; intuitively prompting may work especially well on tasks that the model assigns high probability to observing. If so, this would make prompting more disanalogous in our setup than finetuning. We hope to test this conjecture in future work.

In Appendix D.1, we show a proof of concept that weak-to-strong generalization can still elicit latent capabilities that were never explicitly observed during pretraining, and even when prompting is not possible. In particular, we use AlexNet (Krizhevsky et al., 2012) to supervise models pretrained with DINO (Caron et al., 2021), a self-supervised method in computer vision that learns strong representations. We find that the strong student generalizes significantly beyond AlexNet's performance, even though the student never observed any classification labels during pretraining. Future work should study and mitigate this pretraining leakage disanology more systematically.

6.2 Future Work

What would convince us that we have a "solution" to superalignment? This is a complicated question and we do not claim to have a complete answer. However, we expect substantial progress in at least the following three areas will be necessary: analogous setups, scalable methods, and strong scientific understanding. We now sketch out concrete problems for each of these areas.

6.2.1 Concrete Problems: Analogous Setups

Having strong measurements and a reliable methodology is extremely important for making empirical progress in any field. In particular, it is important that we have metrics which provide strong signal about whether we are making real progress toward the problem we ultimately care about. Important directions for follow-up work include:

- Making our setup more analogous by fixing the main remaining disanalogies described in Section 6.1. Analogous setups are essential to ensure that methods that work today will continue to work for superhuman models.
- Validating that disanalogies are not severe, for example by checking that results are qualitatively similar to using e.g. 3rd grade humans to supervise our strongest models today.
- Relaxing some of the simplifications we made, e.g. by generalizing our methods and results to complicated generative tasks.
- Testing how robust our weak-to-strong classifiers are to optimization pressure when we attain high PGR; for example, if we attain good weak-to-strong generalization with RMs, can we optimize the learned RM using RL?

- Testing our conjecture that prompting-based methods in our current setup will not be as indicative of future results relative to finetuning-based methods (Section 5.2.1), and improvig our setup to fix this.
- Identifying new or more specific disanalogies with our setup and fixing them.

Additionally, we do not yet know what future models will look like. We should update our setup over time as we learn more about how broadly superhuman models will be built.

6.2.2 Concrete Problems: Scalable Methods

One intuition for why major progress on weak-to-strong generalization seems possible is because all we need to do is extract everything the strong model already "knows" about the task of interest—the strong model should intuitively already understand the task, and should hopefully have salient representations of that task. This suggests a number of properties that should be satisfied by the desired generalization, and which we may be able to measure without access to ground truth.

- The desired generalization should be able to *disagree with the weak supervision* when the weak supervision is wrong. This is a property our auxiliary confidence loss may capture.
- The desired generalization should be "*natural*" or "*salient*" to the model. For example, we should not need to change the model too much to elicit the desired concept.
- The desired generalization should be *consistent*. Consistency properties range anywhere from basic logical consistency to complicated forms of consistency between many prompts (e.g. cycle consistency, cross examination, etc.).

Future work should identify additional unsupervised properties that can be used to specify the desired generalization. More generally, there are very likely existing methods in the machine learning literature (e.g. in semi-supervised learning or robust finetuning), which would be natural to try and which could also lead to substantial gains in weak-to-strong generalization. Generalization-based approaches to weak-to-strong learning are complementary to scalable oversight methods, in which the weak supervisor interacts with the strong model to improve the quality of the weak supervision.

6.2.3 Concrete Problems: Scientific Understanding

We will need an extremely high degree of trust and reliability in our methods for aligning superhuman models in high-stakes settings. We will not get this from strong benchmark performance alone. Instead, we also need a thorough understanding of precisely *when* and *why* our methods work. Example questions of interest include:

- What explains the difference between the relatively strong results on NLP datasets and the relatively poor results with reward models when using naive finetuning?
- What makes a concept easy or hard to elicit? What is a good definition of "salience"?
- Can we reliably estimate generalization error at test time without any labels? For example, can we measure the degree of weak-to-strong underspecification (Lee et al., 2022b)?
- Can we reliably extrapolate generalization error across many orders of magnitude using scaling laws?
- How important are the errors in the weak supervision, precisely? How do different kinds of weak label biases affect generalization?
- How robust are our proposed methods to optimization pressure?

In Section 5 we only scratched the surface for understanding weak-to-strong generalization, but future work will need to go much further. An advantage of our setup is that it makes it easy to run simple experiments to scientifically study generalization phenomena across a wide range of settings.

6.3 Conclusion

Recent progress in AI has been faster than almost anyone anticipated (Steinhardt, 2022; Bengio et al., 2023). For an increasing number of researchers, the possibility of superhuman models being

developed this decade has become increasingly plausible. Broadly superhuman models would be extraordinarily powerful and, if misused or misaligned with humans values, could potentially cause catastrophic harm (CAIS, 2022). Given the stakes, we need to establish extremely high reliability in the alignment of these systems ahead of time. But for years it has been unclear how to empirically study superhuman model alignment. We believe it is now easier to make progress on this problem than ever before.

7 Acknowledgements

We would like to thank Boaz Barak, Paul Christiano, Jacob Steinhardt, Ananya Kumar, Jakub Pachocki, John Schulman, Wojciech Zaremba, Alec Radford, Nat McAleese, and William Saunders for valuable technical insights and discussions. We are grateful to Mia Glaese, Boaz Barak, Kush Bhatia, Jean-Stanislas Denain, Erik Jones, Polina Kirichenko, Daniel Kokotajlo, Yoonho Lee, Jessy Lin, Richard Ngo, John Schulman, Peter Tong, Fred Zhang, Ruiqi Zhong, Ryan Greenblatt, Fabien Roger, Paul Christiano, Steven Adler, Rai Pokorny, Adam Kalai, Jacob Hilton, Roger Grosse, Dan Hendrycks, Alec Radford, and Scott Aaronson for helpful feedback on earlier drafts of this paper. We also thank Shantanu Jain, Avital Oliver, Suchir Balaji, Cathy Yeh, and the Platform team for infrastructure help. CB is also grateful to Dan Hendrycks, Jacob Steinhardt, and Paul Christiano for many formative discussions over the years.

References

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pp. 312–321. PMLR, 2019. (Cited on page 33)

Christopher G Atkeson and Stefan Schaal. Robot learning from demonstration. In *ICML*, volume 97, pp. 12–20. Citeseer, 1997. (Cited on page 5)

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. Exploring the landscape of distributional robustness for question answering models. *arXiv preprint arXiv:2210.12517*, 2022. (Cited on page 4)

Stephen H Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pp. 273–282. PMLR, 2017. (Cited on page 4)

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a. (Cited on page 1, 5)

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b. (Cited on page 5, 47)

Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Machine Intelligence 15*, pp. 103–129, 1995. (Cited on page 5)

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. (Cited on page 4)

Qiang Ning Ben Zhou, Daniel Khashabi and Dan Roth. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *EMNLP*, 2019. (Cited on page 29)

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023. (Cited on page 18)

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. (Cited on page 4)

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10925–10934, 2022. (Cited on page 4)

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *OpenAI Blog*, 2023. (Cited on page 47)

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020. (Cited on page 29)

Sam Bowman. Artificial sandwiching: When can we test scalable alignment protocols without humans? *AI Alignment Forum*, 2022. (Cited on page 5)

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022. (Cited on page 2, 47)

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. (Cited on page 14)

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. (Cited on page 5)

CAIS. Statement on ai risk, 2022. (Cited on page 4, 19, 47)

Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power? *arXiv preprint arXiv:2311.08379*, 2023. (Cited on page 48)

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. (Cited on page 17, 35, 40)

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021. (Cited on page 36)

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020a. (Cited on page 35)

Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. *Advances in Neural Information Processing Systems*, 33:21061–21071, 2020b. (Cited on page 33)

Paul Christiano. Approval-directed bootstrapping. *AI Alignment Forum*, 2018. (Cited on page 8)

Paul Christiano. Capability amplification. *AI Alignment Forum*, 2019. (Cited on page 8)

Paul Christiano, Buck Shlegeris, and Dario Amodei. Supervising strong learners by amplifying weak experts. *arXiv preprint arXiv:1810.08575*, 2018. (Cited on page 2, 5)

Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge. Technical report, Alignment Research Center (ARC), 2022. (Cited on page 5, 11, 17, 44)