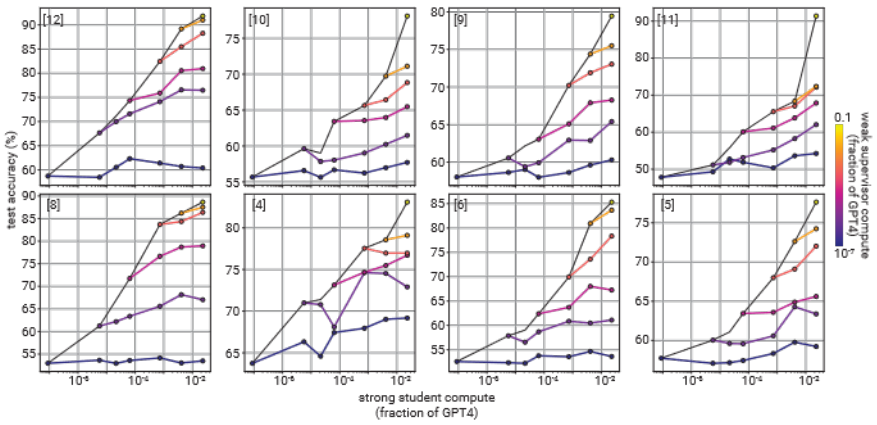

Figure 23: **Linear probing qualitatively matches finetuning weak-to-strong generalization.** Test accuracy as a function of strong student compute on a subset of our NLP tasks. Inset numbers indicate dataset id (compare Figure 12). Accuracy of a linear probe on student model trained with ground truth in black, accuracy of linear probe on students trained directly with weak linear probe supervision shown in solid lines with circles (hue indicates compute of weak supervision).

D.2 Linear probing

In addition to our main finetuning experiments, we also perform weak-to-strong generalization experiments in the linear probing setting. We freeze all weak and strong model parameters, and train new linear classification heads both using ground truth labels and using weak labels. We train linear probes with Adam optimizer (Kingma & Ba, 2014), 10^{-3} learning rate, batch size 128, and no weight decay for 200 epochs, for both weak and strong model training. We do early stopping based on agreement to the weak labels on the validation set and report test accuracy. Results are shown in Figure 23. We observe qualitatively similar generalization compared to the full finetuning case.

Generally, we found the linear probing setting to be very useful to quickly iterate on methods, datasets and ideas. While finetuning provides better results, the qualitative trends in linear probing are similar, and the experiments are much faster and easier to run. For example, we initially found positive results with confidence loss (Section 4.3) and bootstrapping (Section 4.3.1) in the linear probing setting.

E The effects of weak label structure

One challenge in weak-to-strong generalization is the presence of errors in the weak labels. Throughout most of this paper, we consider a particular type of weak error structure: the kinds of errors smaller, capacity-constrained language models make. However, this is not the only type of errors possible.

In this section, we analyze synthetic examples of other kinds of weak label structures, and the implications they have on generalization. Weak model error structure must be considered in relation to the particular strong model at hand. For example, we conjecture that the extent to which the strong model can imitate the weak supervisor may be very important. If we have two strong models of the same performance on the actual task but one is very good at imitating the labels, then we expect that model will generalize less desirably, at least with the naive finetuning method.

In Section 5.1.3 we found that surprisingly the strongest students are imitating the weak supervisor mistakes less than smaller student models in our setting. Since we expect superhuman models to be very good at imitating human supervisor, this may be a major disanalogy. In this section we test cases where the weak supervisor can be imitated easily.

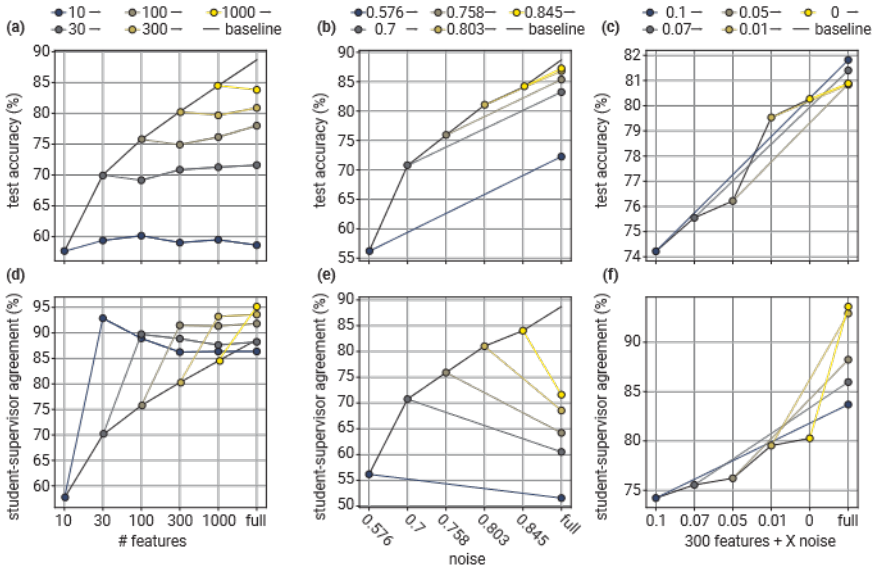

Figure 24: **Synthetic experiment on simulation difficulty.** We consider three types of weak errors in a linear probing setting: **(a,d)** perfectly simulatable, where weak models use a subset of strong model features; **(b,e)** completely unsimulatable, where the weak labels are obtained by applying random noise to the ground truth; **(c,f)** a mixture of the two settings, where label noise is applied to perfectly simulatable weak labels. Top row of panels shows test accuracy and bottom row shows agreement to the weak labels. In addition to weak label accuracy, the structure of mistakes plays a major role in weak-to-strong generalization.

E.1 Synthetic experiments on simulation difficulty

First, we consider a simplified linear probing setting, where we can ensure that the student can perfectly simulate the supervisor predictions by construction. Specifically, we extract a representation $X \in \mathbb{R}^{n \times d}$ of the SciQ dataset using a model of an intermediate size in the GPT-4 family, where n is the number of datapints, and d is the dimensionality of the residual stream (Elhage et al., 2021). We can then consider the family of linear models[10] $\mathcal{M}_k$ where $k \leq d$ by training a linear probe only on the first k features extracted by the model. In particular, for $k = d$ we recover the standard linear probe. By construction for $k_1 \geq k_2$, the model $\mathcal{M}_{k_1}$ can perfectly simulate $\mathcal{M}_{k_2}$.

Next, we can run our standard weak-to-strong generalization experiment, following the setup described in Section 3, using the family of models $\mathcal{M}_k$. We train the weak supervisor models on $10k$ datapoints, and produce hard weak labels on the remaining $13k$ datapoints. We report the results in Figure 24(a,d). In this setting, the simulation is very easy, and we do not observe substantial improvements in the strong student model compared to the supervisor performance. The test agreement values are substantially higher than the weak model accuracy, indicating that the students are overfitting to the supervisor errors. Interestingly, even in this simple setting the agreements are not 100%, likely due to the fact that the student models are trained on finite data, and with light l_2-regularization.

We can also consider the opposite setting: what if the student model cannot simulate the mistakes of the weak teacher at all? Specifically, we generate weak labels by randomly flipping the labels to match the accuracy of the weak models from the previous experiment. As a result, we get weak labels with the same accuracy, but which are completely unpredictable. In Figure 24(b,e), when we train the student model on the these weak labels, we can get substantially higher accuracy than the accuracy of the weak labels. In other words, if the errors of the weak supervisor are completely unpredictable (random) for the student, with enough data we should be able to recover good generalization, substantially exceeding the performance of the supervisor.

[10]We train logistic regression using the default parameters in the `sklearn.linear_model.LogisticRegression` class (Pedregosa et al., 2011) for this experiment.

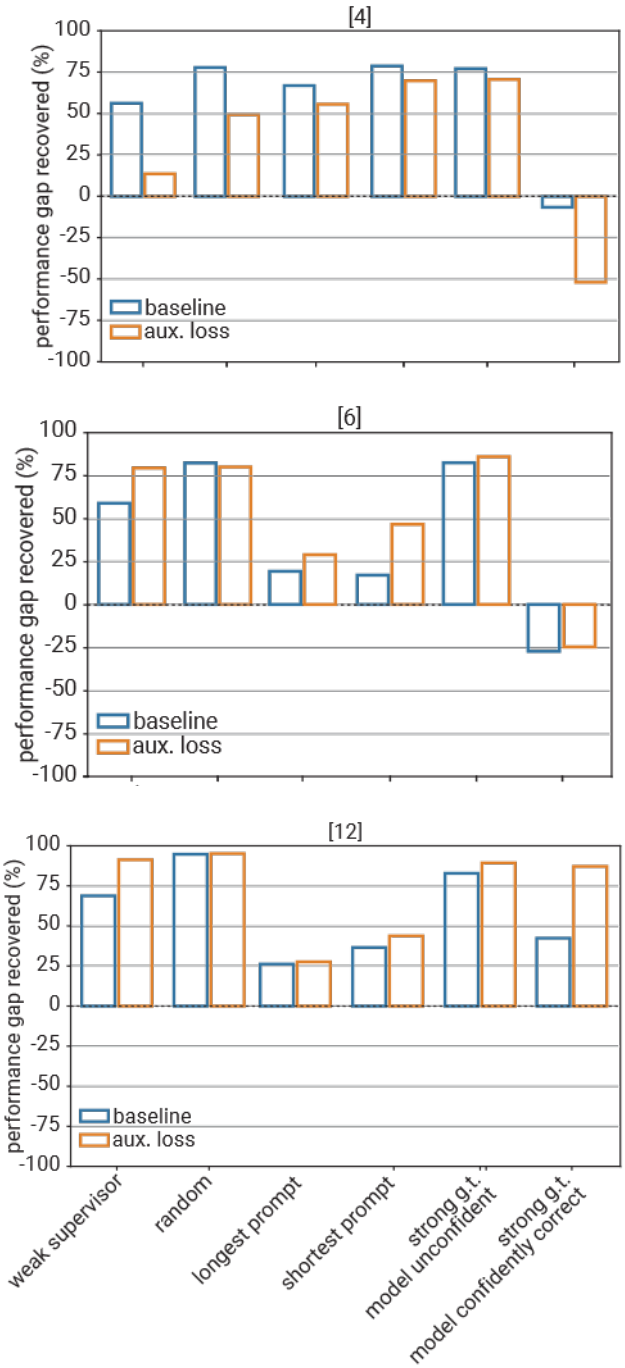

Figure 25: **PGR for weak labels with same accuracy but different error structures.** The inset number in each panel indicates the dataset (compare Figure 12). Weak-to-strong generalization and methods both depend critically on the structure of the weak supervisor errors. While it is trivial to pick error structures that generalize well (for instance, random noise), these error structures are also very disanalogous to the ultimate superalignment setting, where we want to study the structures of human errors.

Finally, in Figure 24(c,f) we consider a mixture of these two settings: we start with a perfectly simulatable weak model $\mathcal{M}_{300}$, and then add various amounts of label noise to the resulting weak labels. By training a strong student model (using all features) on the resulting weak labels, we recover the performance close to the performance of $\mathcal{M}_{300}$.

Discussion of results. The simple experiment in this section suggests that in addition to the weak label accuracy, it is important to consider the *structure of weak errors*. In particular, if the weak errors are extremely easy for the strong model to simulate, the student may not generalize much better than the weak supervisor with naive finetuning on the weak labels. On the other hand, if the mistakes of the weak supervisor are completely unpredictable, the student can denoise the predictions of the supervisor and generalize better. In future work, we believe it is important to consider various types of weak supervision with different structures of mistakes, and build a better understanding of how they affect weak-to-strong generalization.

E.2 Different weak error structure means different generalization

To further explore the impact of different weak error structures, we created several synthetic sets of weak labels for each dataset, all with error rate identical to the weak model's error rate. To construct these labels, we start from ground truth, and then flip a subset of labels to match the accuracy of a particular weak model. We target a few types of error structures, such as pure noise, easy-to-model bias, hard-to-model bias, and adversarial bias.

In particular, we looked at:

1. `weak supervisor`: the baseline — labels are generated in the same way as in the rest of the paper
2. `random`: flip the label of random datapoints
3. `longest prompt`: flip the label of longest datapoints by characters
4. `shortest prompt`: flip the label of shortest datapoints by characters
5. `strong g.t. model unconfident`: flip the label of the datapoints that the strong ceiling model is most unconfident on
6. `strong g.t. model confidently correct`: flips the label of the datapoints that the strong ceiling model is most confidently correct on

Despite all of these weak labelers having the same weak accuracy, we find that the generalization can vary wildly depending on the structure of the weak errors. We report the results in Figure 25.

Furthermore, the dynamics of supervisor-student agreement through training can have qualitatively different behavior (Figure 26). For errors coming from a weak model, we see that there is often initially a period of generalization, followed by a period of overfitting where it learns the weak model's errors. The confidence auxiliary loss mitigates this overfitting. For easy-to-fit error structures such as `longest prompt`, the overfitting happens much faster. For other kinds of errors, such as random noise, we often see that generalization improves throughout: weak errors are not modeled, but the signal from the weak model is.

E.3 Making imitation trivial

One possible major disanalogy in our setup, as discussed in Section 6.1, is the fact that our models are not very good at imitating the weak model[11] (Section 5.1.3), but superhuman models may be very good at imitating humans. It is possible that if the strong model were good at imitating the weak model, then it would generalize substantially less desirably by default.

To test an extreme version of this hypothesis, we create a synthetic setting where the strong model can trivially imitate the weak model very well. In particular, we modify the task by appending "I think this is {`weak_label`}. What do you think?" to every prompt, where weak_label is "correct" or "incorrect" based on the weak model prediction. In this case, the hardened weak label is present in-context, and the simulation is trivial.

[11] Also known as learning the "human simulator" in the terminology of Christiano et al. (2022).

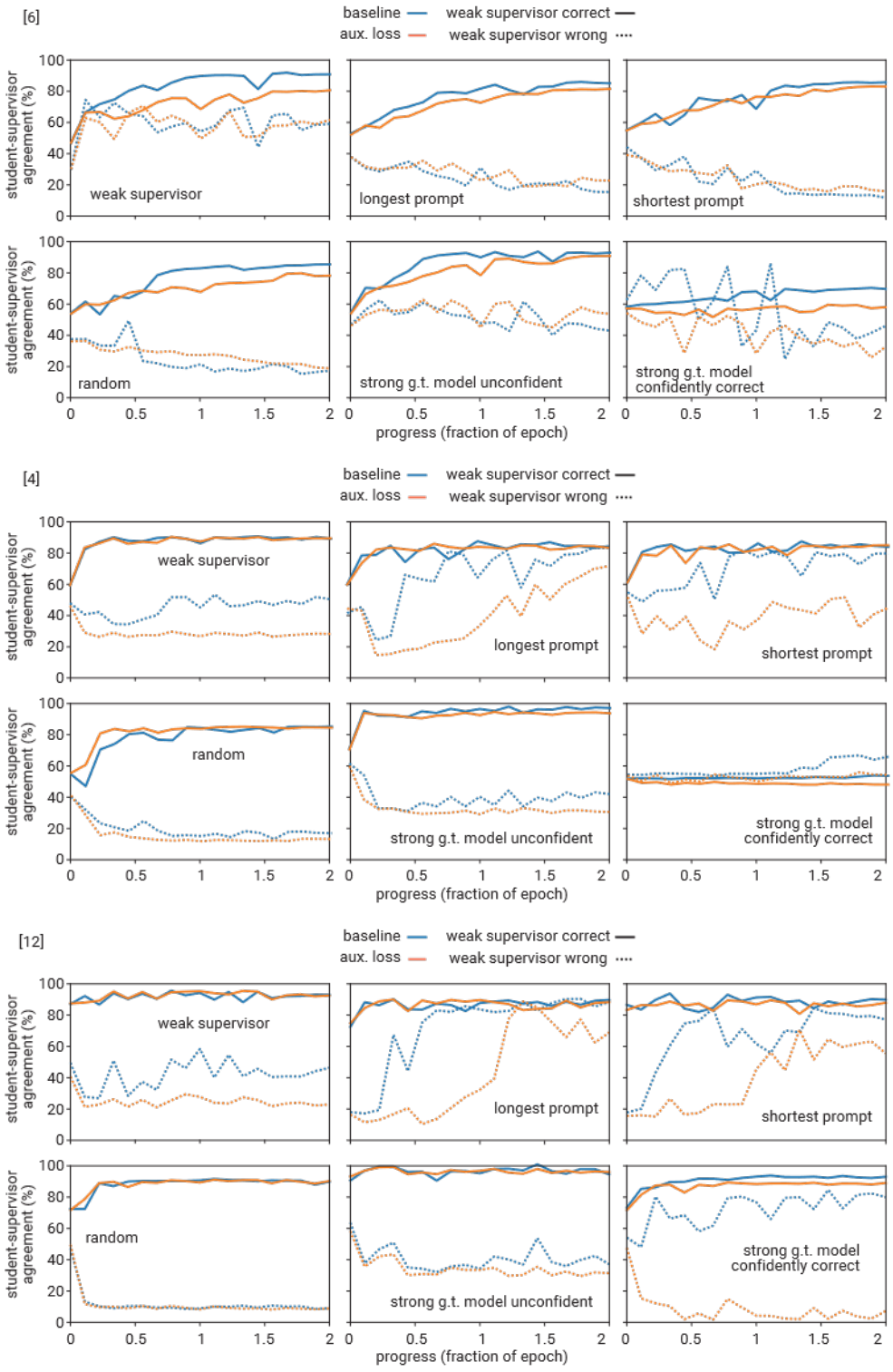

Figure 26: **Training dynamics change for different weak errors.** We show teacher-student agreement for different weak error structures on three datasets. We see that the training dynamics have qualitatively different behavior for different error structures, despite all weak labelers having the same accuracy.

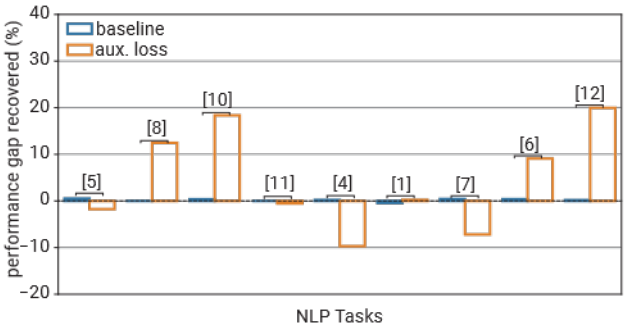

Figure 27: **Generalization when emulating weak labels is trivial.** Very little weak-to-strong generalization occurs if emulating the weak labels is trivial: average PGR across tasks is 0.002 ± 0.003 for baseline, and 0.046 ± 0.108 for aux loss, compared to around 0.2 and 0.8 respectively for the original tasks.

As expected, we find that both the baseline and the confidence loss introduced in Section 4.3 show poor weak-to-strong generalization (Figure 27) in most cases. Interestingly, the confidence loss still improves upon the baseline achieving non-trivial generalization in several tasks.

F How should we empirically study superalignment, methodologically?

What makes a setup good for studying superalignment in the first place, all things considered? Tractability and ease of study are clearly important criteria, but also certainly not the only ones. This question is non-obvious because superalignment is qualitatively different from other machine learning problems: it is a problem we will face in the future, not a problem that we face today. Nevertheless, it is crucial that we solve this problem *before* it becomes serious, as even a single failure of superintelligence misalignment in practice could be catastrophic.

This presents a major methodological challenge: how do we even approach studying a problem that is not yet a problem? How do we make progress on the core difficulties of superalignment? How do we make progress with today's systems, knowing that our efforts will not be wasted by surprising new model capabilities that will inevitably arise in the future (Wei et al., 2022)? We do not claim to have a complete answer to these questions, but we outline some best practices for maximizing our chances of making real progress on superalignment.

Analogous setups. We should construct increasingly analogous empirical setups, and we should enumerate any remaining disanalogies. A setup is analogous if our results on that setup do not rely on assumptions that will break down in the future, making results today likely qualitatively similar to results in the future. Our main evaluation setup, introduced in Section 3, is intended to be more analogous to the superalignment problem. We enumerate some remaining disanalogies with our setup in Section 6.1.

Enumerating assumptions. We should enumerate the key assumptions that our results (either implicitly or explicitly) rely on. Clarifying what assumptions we are making makes it much easier to know when our results might break down. We enumerate our main disanalogies and assumptions in Section 6.1 and Appendix G.3.

Sensitivity analysis. We should evaluate the sensitivity of our results to changes in our assumptions and empirical setup. While we can make informed guesses about the future, we do not know exactly what future models will be like, so it is difficult to entirely trust any particular experimental setup. Validating that our results are robust to many different sets of assumptions can make us substantially more confident our results will transfer to the future superalignment problem. We do some initial sensitivity analysis in Appendix E, and intend to do much more in future work.

Scalable techniques. We should avoid techniques that rely on assumptions that will likely break down for future (superhuman) models. For example, when we do few-shot prompting we are in-

tuitively incentivizing models to predict some useful distribution of human text, whereas when we do finetuning we are intuitively incentivizing a model to output what it knows regardless of how it knows it. This is one of the reasons we focus on finetuning methods in this paper: they are more likely to scale to superhuman models compared to prompting.

Incidental usefulness today. One possible validation that progress on our setup is real would be to show that it is incidentally useful in practice today; while we advocate focusing on the core challenges of superalignment, if our findings are never useful with today's models that would be evidence that we are not on the right track. One example of a near-term practical milestone would be to align GPT-4 on instruction-following tasks using only GPT-3-level supervision; if we could get strong alignment without any humans involved at all, that would make alignment much simpler and cheaper today. However, usefulness today is certainly not sufficient for aligning superintelligence, and in general a common failure mode of empirical alignment research is it prioritizes usefulness today at the expense of analogousness and scalability.

Updating over time. We should update our evaluations and validate past findings as we learn more about what future models will look like. While we focus on the pretrained language model paradigm today, we plan on updating our setup if or when this stops being the dominant paradigm.

G How weak-to-strong generalization fits into alignment

Superintelligent AI systems will be extraordinarily powerful; humans could face catastrophic risks including even extinction (CAIS, 2022) if those systems are misaligned or misused. It is important for AI developers to have a plan for aligning superhuman models ahead of time—before they have the potential to cause irreparable harm.

Our plan for aligning superintelligence is a work in progress, but we believe that weak-to-strong techniques could serve as a key ingredient. In this section we sketch several illustrative possiblities for how we could use weak-to-strong generalization to help align superintelligent systems.

G.1 High-level plan

Leike & Sutskever (2023) propose the following high level plan, which we adopt:

1. Once we have a model that is capable enough that it can automate machine learning—and in particular alignment—research, our goal will be to align that model well enough that it can safely and productively automate alignment research.

2. We will align this model using our most scalable techniques available, e.g. RLHF (Christiano et al., 2017; Ouyang et al., 2022), constitutional AI (Bai et al., 2022b), scalable oversight (Saunders et al., 2022; Bowman et al., 2022), adversarial training, or—the focus of this paper—-weak-to-strong generalization techniques.

3. We will validate that the resulting model is aligned using our best evaluation tools available, e.g. red-teaming (Perez et al., 2022a;b) and interpretability (Ribeiro et al., 2016; Olah et al., 2018; Bills et al., 2023; Li et al., 2023).

4. Using a large amount of compute, we will have the resulting model conduct research to align vastly smarter superhuman systems. We will bootstrap from here to align arbitrarily more capable systems.

The goal of weak-to-strong generalization is to ensure step (2) is solved: align the first model capable of automating machine learning and alignment research. Importantly, this first model will likely be qualitatively superhuman along important dimensions, so RLHF is unlikely to be sufficient (Section 4). If we had a superhuman model, how would we apply weak-to-strong generalization to align it?

G.2 Eliciting key alignment-relevant capabilities with weak-to-strong generalization

There are many different alignment-relevant capabilities we could try to elicit from a superhuman model that could significantly help with alignment, including:[12]

- **Safety:** does a given behavior produced by an AI system risk the safety of human lives or well-being in important ways?
- **Honesty:** is a given natural language statement true or false?
- **Instruction following:** does a given behavior produced by an AI system follow a user's instruction faithfully?
- **Code security:** does some given code have important security vulnerabilities or backdoors? Is it safe to execute it?

In the ideal case, the capability we elicit from the model would be robust enough that we can turn it into a reward model and safely optimize it; future work should assess the feasibility of this approach. At the opposite extreme, we could potentially use the elicited capability as an "oracle" that we can manually query; intuitively, if we had a superhuman oracle model, we may be able to leverage it to help us bootstrap to a more robust alignment solution, even if that oracle is not itself entirely robust.

G.3 Alignment plan assumptions

Many alignment plans which appear different on the surface actually depend on heavily correlated assumptions. For a given alignment plan, it is also often unclear which subproblems the plan attempts to solve, and which subproblems the plan assumes are unlikely to be an obstacle. As a result, we think enumerating assumptions is an important part of making progress on alignment.

In addition to the major disanalogies discussed in Section 6.1, the assumptions we make for an alignment plan based on weak-to-strong generalization include:

- **No deceptive alignment in base models.** We assume that pretrained base models (or the equivalent in future paradigms) will be highly intelligent but not highly agentic (e.g. will not have long-term goals)—and consequently will not be deceptively aligned (Hubinger et al., 2019; Ngo et al., 2022; Carlsmith, 2023) out-of-the-box. Our goal is to elicit the superhuman capabilities of this capable but safe base model, and use those capabilities to create an aligned (possibly agentic) superhuman model.
- **Elicited concepts are sufficiently robust, or do not need to be.** We assume it is either possible to solve alignment using only a small amount of optimization applied to the capabilities we elicit, or that it is possible to make weak-to-strong elicited capabilities sufficiently robust against overoptimization.
- **The concepts we care about are natural to future AGI.** The superhuman base model we apply weak-to-strong generalization to has some "alignment-complete" concept, such as honesty, that is extrapolated in the way we would endorse if we could understand everything the superhuman model understands, and which is natural enough to the model that it is feasible to elicit.
- **Sufficiently gradual takeoff.** Before we have superintelligence, we will have somewhat superhuman models long enough that we can use them to finish solving the full superintelligence alignment problem. We can use it to solve superalignment before it causes recursive self improvement or catastrophic damage.
- **Moderately superhuman models are sufficient to solve alignment.** We assume the first models capable of automating alignment research in practice are moderately superhuman, i.e. in a regime similar to what we study empirically in this work. For example, we may assume that we only need to bridge a weak-strong gap of at most (say) 4 OOMs of effective compute.

[12] Ideally we elicit several related concepts and verify that we get consistent answers between them.

- **No need to solve human values.** We assume we do not need to solve hard philosophical questions of human values and value aggregation before we can align a superhuman researcher model well enough that it avoids egregiously catastrophic outcomes.

This list represents a non-exhaustive set of notable assumptions we often operate under, and we will constantly reassess and update these assumptions over time as we learn more. We *do not* think these are necessarily valid assumptions by default, and believe it is important to validate them, work towards making them true, or mitigate failure modes from them being invalid.

Furthermore, there are a huge number of uncertainties about what future AI systems will look like and exactly how we should align them.