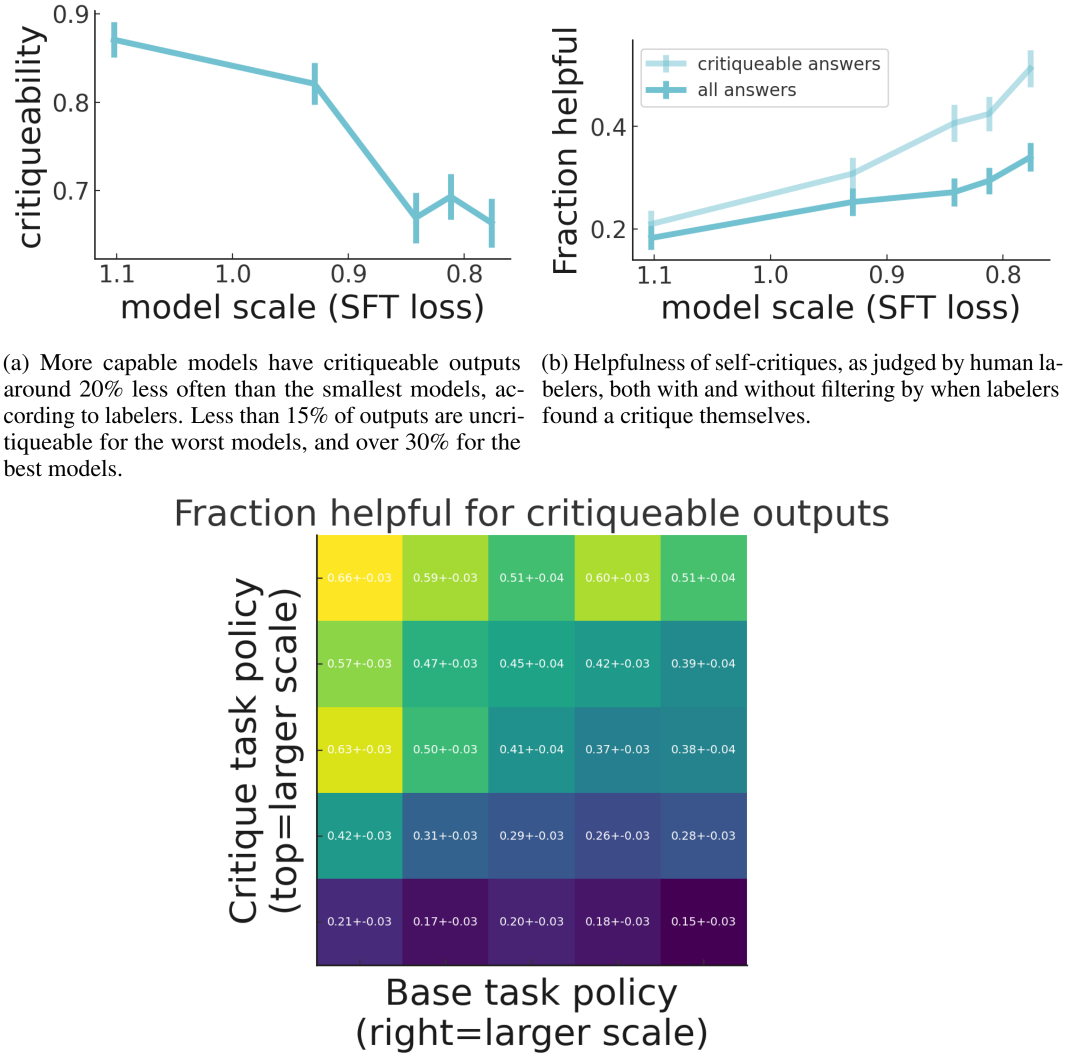

(a) More capable models have critiqueable outputs around 20% less often than the smallest models, according to labelers. Less than 15% of outputs are uncritiqueable for the worst models, and over 30% for the best models.

(b) Helpfulness of self-critiques, as judged by human labelers, both with and without filtering by when labelers found a critique themselves.

(c) Larger models are not only better at critiquing, but harder to critique – even filtering for only cases where labelers found a critique. The diagonal (spanning lower left to upper right) corresponds to the “critiqueable answers” line in 4b.

Figure 4: More capable models are significantly better at self-critiquing (Figure 4b). Although more capable models get better at generating hard-to-critique answers (Figure 4c), their ability to critique their answers is improving more rapidly with scale. This is true even without adjusting for the fact that humans find fewer critiques of more capable models (Figure 4a). In all figures, we sample at the same random temperature for both the base task and critique task; the effects are equally visible at all temperature ranges (not pictured).

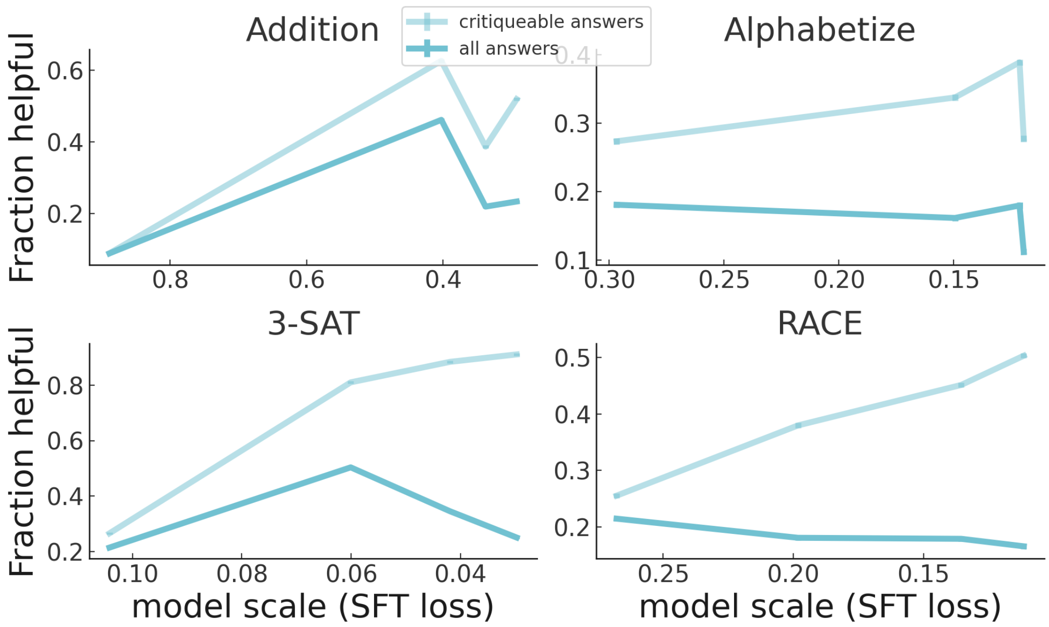

Figure 5: Helpfulness of self-critiques for synthetic tasks, according to a critique validity oracle. Like Figure 4, we show with and without filtering for critiqueable answers (according to a critiqueability oracle).

On topic-based summarization, we find that larger models are better at critiquing themselves (Figure 4b), even without filtering for critiqueable answers. This holds even though answers from larger models are harder to critique (Figure 4a, 4c).

One caveat is that our supervised dataset contains more critiques of outputs from larger models, since we typically use relatively capable answer models. However, we believe this effect to be minimal.

On synthetic tasks, we generally observe similar trends in the critiqueable case (Figure 5), though the story is less clear. Overall, we have no strong reason to believe positive critique scaling to be a fundamental trend. We also do not know, for example, whether the trend would also go away if we use reinforcement learning to train both the answer and critique model. Nevertheless, we believe models have only recently reached a scale where critiquing on realistic tasks is possible.

4.3 Refinements

Another check of whether model-generated critiques are useful is to compare critique-conditional refinements to direct refinements. In other words, we compare refinements generated using only an answer to refinements generated using both an answer and a critique of that answer.

In order to improve conditional refinement performance, we can improve the critique. To do that, we do best-of-N [SOW+20] against the helpfulness score; we sample N critiques, choose the best according to the model's helpfulness score, and use that critique for the conditional refinement. For direct refinements, we take best-of-N refinements using our model's critiqueability score.

In our refinement experiments we ask for a refinement regardless of whether the initial answer is critiqueable. If the initial answer were perfect, the model would have no chance at improving it. Thus in order to not "force" the model to refine, we compare the refinement to the original using the model's critiqueability score.

We also include baselines of the original "best-of-1" sample, and a best-of-8 sample (generating new answers from scratch, and ranking them by critiqueability). These experiments use temperature 0.5 to sample, which we believe to be near optimal for best-of-1 on all tasks (answering, critiquing, and refinements).

4.3.1 Findings

Our results are depicted in Figures 6 and 7 and samples can be found in Appendix F.3. Despite being somewhat noisy, these results suggest:

1. **Good critiques help refinement.** Good critiques are useful for refinement. Conditional refinement appear to outperform direct refinements, but only with critiques selected via best-of-N against helpfulness. Larger N helps improve the conditional refinements.
2. **Large model scale enables refinements.** Both forms of refinement significantly outperform the original output for larger models, but have little to no effect for smaller models.
3. **Using critiques may not be competitive if controlling for compute.** Rejection sampling to select better critiques to use for refinements is competitive with rejection sampling on answers, a roughly compute-equalized baseline.[5] However, rejection sampling on direct refinements appears to be a stronger baseline.

5 Generator-discriminator-critique (GDC) gaps

In this section, we present results suggesting that models are not articulating all the problems they “know about.” Furthermore, despite the positive results in critique scaling from Section 4.2, we do not see evidence that the gap between our models’ discrimination and critique writing abilities is closing with scale.

5.1 Setup

In this section we consider the following three tasks:

- G: answer generation
- D: answer discrimination (critiqueability)
- C: answer critiquing

In our main results from Section 4.2, we compared tasks G and C: To what extent can a model critique its own answers when they are poor? Comparing G and D is also interesting: Can a model tell when its own outputs are good or poor? As with critique scaling, we have two competing trends: The discriminators are getting better in an absolute sense, but the critiqueable answers may also be getting harder or subtler to critique.

Finally, we argue that the gap between D and C is especially interesting: if a model can tell an answer is poor, can it also point out the flaw to a human? If we could train models to always point out when they notice flaws, this could go a long way towards having trustworthy and aligned models. For more discussion, see Appendix C.

This motivates us to measure these quantities in such a way that:

- The different tasks can be compared on the same axis. For each pair, we will aim to measure a "XY gap" measuring the amount Y performance exceeds X performance
- The GC gap corresponds to effectiveness of self-critiquing. A positive gap corresponds to ability to improve or check outputs by showing humans critiques.
- The GD gap corresponds to the model’s ability to know when answers it produces are poor. A positive gap corresponds to ability to improve outputs using a discriminator.
- The CD gap corresponds to the model’s ability to give human-understandable critiques on answers it “knows” are flawed (and *inability* to give convincing critiques on sound answers).

Our hope is to ultimately use critiques for better training signal on difficult tasks. In a sense, we would like to take measurements that let us scope out how well this works without actually training our models on this task (see Appendix C.3.3).

[5] This is mildly surprising since rejection sampling on answers gives "fresh starts" while refinements are sometimes forced to start with a poor answer. We speculate that with enough compute budget, it is optimal to use a combination of the two, as well as iterative refinement.