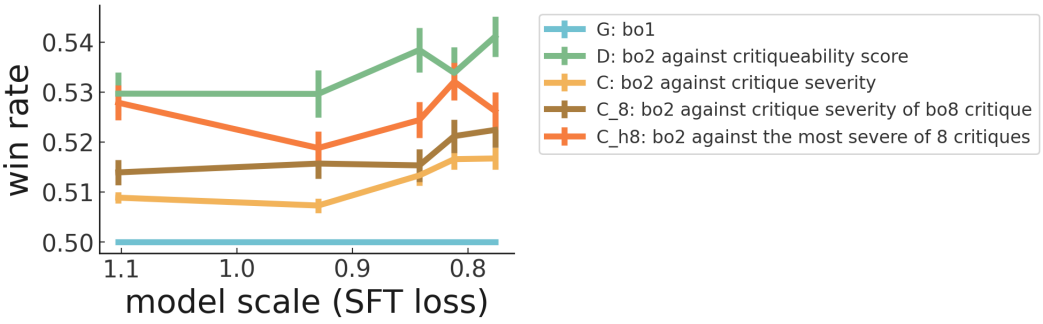

Figure 10: Variants of C with humans rating severity of multiple critiques (C_{h8}) or rating severity of a single critique optimized via best-of-8 against a helpfulness model (C_8). Both versions outperform the basic version with just one critique per answer. Unsurprisingly, humans evaluating 8 critiques outperforms humans evaluating 1 optimized critique.

C.3 Measuring gaps discussion

Recall that we proposed the following way of measuring GDC gaps in Section 5:

- G: What is the average performance of a generator sample?
- D: What is the performance of the generator with best-of-N against the discriminator?
- C: What is the performance of the generator with best-of-N against the severity of a critique?

C.3.1 Reasons to expect negative CD gaps

With our above definition, C does not only measure a model's ability to articulate critiques. It also uses a human to check the critique validity, thus letting C directly search against an oracle. Thus, because our D (critiqueability) models do not match the labels they are trained on, we see negative CD gaps on simple tasks (See Figure 9).

To make C more analogous to D, we could take the definition of C one step further and train a model to predict critique validity and severity (or preference). In other words, the model should not only be able to produce a good critique, but it should also "know" that it is good. Since we did not train validity/severity models, we instead use the helpfulness model, which gives:

- C_m: What is the performance of the generator with best-of-N against the helpfulness score of a critique? That is, we use the helpfulness model as a discriminator: to judge an answer we generate a critique and consider the answer poor if the helpfulness score is high.

Note also that we expect no difference between C_m and C in the limit of very large models, since learning human helpfulness labels should become easy. Even then, we could expect $D \leq C$, if:

1. Vocalizing critiques helps a model understand how to discriminate, as a "chain of thought" [WWS+22b, WWS+22a]
2. More generally, if we do not control for compute. For example, we could search for critiques (see Section D)

Recall that in Section 5 we found a negative CD gap for the Addition, Alphabetize, and RACE synthetic tasks. We suspect this is due to C's usage of the oracle and that we would have $C_m < D$ but do not investigate further in this work.

C.3.2 Alternative C definitions

Consider the following variants to C, involving using 8 critiques and various amounts of model versus human involvement:

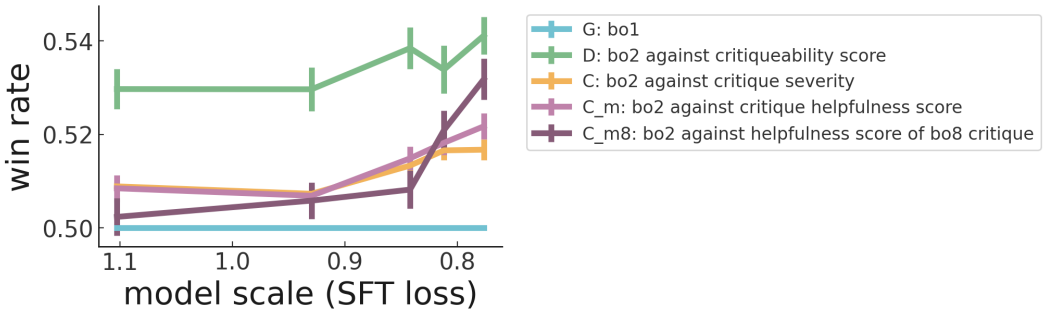

Figure 11: Variants of C with no human - i.e. rather than using a human severity rating, we simply use a helpfulness score. Generating many critiques and taking the best according to helpfulness score seems to improve the helpfulness discriminator for large models. Ideally M $\to \infty$ catches up to or exceeds discriminator performance.

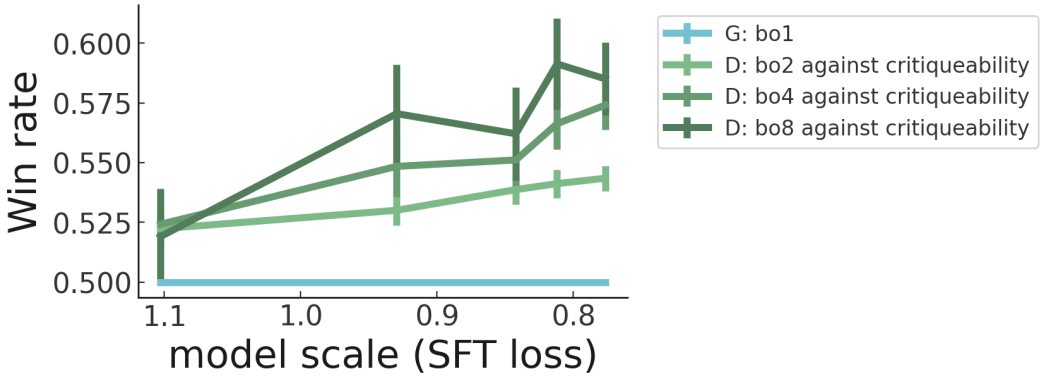

Figure 12: GD gap with increasing N, i.e. win rate of best-of-N against the critiqueability model vs. best-of-1. We generate answers from the same model, and use human rankings as ground truth. Gains from best-of-N seem to improve slightly with model size.

1. C_{h8}: Choose 8 critiques to show to the human, and pick the answer according to validity and severity of the most critical critique. Essentially, we show the human a slate of critiques (like we did in Section 3.4)

2. C_8: Choose a single best-of-8 (according to helpfulness score) critique to show to the human, who uses validity and severity in order to judge the answer. This is just like C, but with a better critique.

3. C_{m8}: Choose a single best-of-8 (according to helpfulness score) critique, and simply use helpfulness score. This cuts out the human from the loop entirely but relies on helpfulness being comparable across different answers (ideally we would use a severity model).

C_{h8} corresponds to giving humans a chance to review multiple pieces of assistance, similar to our assistance experiments in Section 3.4. C_{m8} corresponds to training a helpfulness model and using an optimized critique model to determine critiqueability, similar to basic versions of debate.

Figures 10 show the first two variants. Unsurprisingly, using a best-of-8 critique helps ($C_8 > C$). Also unsurprisingly, humans evaluating 8 critiques outperforms humans evaluating 1 optimized critique ($C_{h8} > C_8$). However, it still seems to fall short of discriminator ability.

Figure 11 shows the model-only variants, where results are still noisy and perhaps more confusing. C_m appears to be competitive with C. However, using more critiques does not seem clearly useful (C_m vs C_{m8}).

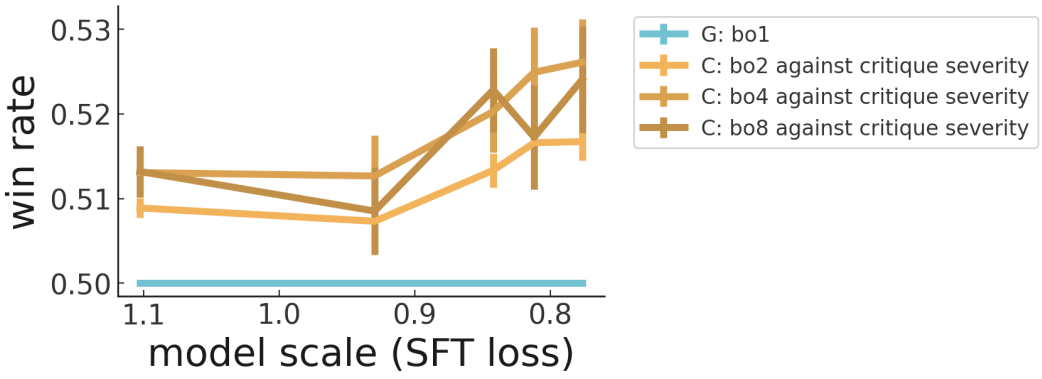

Figure 13: GC gap with increasing N, i.e. win rate of best-of-N against critique helpfulness and severity vs. best-of-1. We generate answers and critiques from the same model, and use human rankings as ground truth. Overall the results suggest our critique models do not make for robust discriminators. Best-of-4 appears consistently better than best-of-2, but best-of-8 possibly does worse than best-of-4 (though noisy). Gains from best-of-N do not appear to improve with model size.

C.3.3 Evaluating reward signal without training

One way to think of the GD gap is that best-of-N checks whether reward learning is working in RLHP, without actually training a model via RL [SOW+20]. (Though we train discriminators for our GDC gaps, it would have been equally sensible to use a preference-based reward model.) The GC gap analogously checks the training signal from using critiques without actually training, if we use a human checking a critique as a discriminator.

Note that with our definitions, GD and GC gaps can only be negative if the discriminator and critique-discriminator, respectively, are worse than chance. One way this can happen is if the generator is over-optimized [MG18] against the discriminator or critique model.

Figure 12 shows GD scaling with N, and Figure 13 shows GC scaling with N. These test in-distribution robustness of our critiqueability score, and robustness of using critiques as a training signal.

D 2-step debate

Our assistance experiments in Section 3.4 serve as a de-risking experiment for 2-step recursive reward modeling: we verify that assisting humans results in better critiqueability labels. If our base task evaluations are better, then we have a better training signal for the base task.

How about debate? A simple version of 2-step debate might look like the following: to evaluate a question, we generate an answer, then generate a critique. A human judges the answer as good if and only if the critique is *not* helpful. We want to compare this judgement to a human judging an answer directly.

Thus, to de-risk debate, we should imagine a critiqueability model trained on flawed labels compared to a critique-severity model trained on labels for critiques from an optimized critique model. Since we don't have a critique severity model, we simply use helpfulness score. We can also use helpfulness score to optimize critiques via best-of-N. Thus overall this simplifies to: compare critiqueability score as a discriminator to "helpfulness score of best-of-N critiques", which is essentially D vs C_m (defined in Appendix C.3.2) but on a different distribution of answers.

We use our dataset of paired misleading and honest answers, since we would like ground truth that does not rely on humans finding critiques. We measure accuracy of picking the honest answer over the misleading answer.

We use our largest model for all tasks and we use temperature 0.5 to sample critiques. We find (see Figure 14):

1. The best-of-N helpfulness score discriminates better with increasing N