Our definition for C is not necessarily canonical, but was chosen to be convenient to measure (see Appendix C.3.2).

With this definition, it is clear that we should generally expect a non-negative GD gap and GC gap. If not, then optimizing against the discriminator or critique training signal makes the generator actively worse. What to expect for the CD gap is less clear. For more discussion, see Appendix C.3.

On a practical level, using best-of-N to measure discrimination ability has many benefits:

- Unlike accuracy, we don't need to calibrate to the distribution. Our critiqueability models are all trained on the same dataset, but answer models have very different critiqueability (Figure 4a).
- We are most interested in the ability to discriminate between answers to the same question, rather than between answers to different questions, which is conflated by the discriminator's ability to tell whether a question is difficult. Though our work uses discriminators, this also means the definitions generalize naturally to using a preference based reward model for D.

For our primary results, we use $N = 2$. This still leaves us with choices for how to measure performance of a sample, and how to measure critique quality.

5.2.1 Sample quality

We explore a number of ways to measure sample quality:

- **Likert**: We ask labelers to rate answers qualitatively on a 1-7 Likert scale.
- **Uncritiqueability (binary label)**: We ask for the fraction of "perfect" outputs according to humans. For synthetic tasks, we use the oracle, so uncritiqueability is the same as accuracy.
- **Win rate**: We ask labelers for comparisons of the sample in question to a sample from some fixed distribution. In particular, if we use win rate against the generator (we can vary the distribution across model sizes), this makes measurements somewhat natural to compare across model sizes. This makes it so that G is always 0.5, and D and C are always at most $1 - \frac{1}{2N}$. In the case where the outcome is binary, win rate is simply a linear transformation from accuracy: $(1 + \text{ACC} - \text{acc}_G)/2$.

A large caveat is that for topic-based summarization, these methods are all potentially flawed, as suggested by our results in Section 3.4. This is especially problematic if labelers fail to find critiques in systematic ways (rather than, e.g. independently failing to find critiques with some probability).

5.2.2 Critique severity

For critique severity measurement, the most straightforward thing is to use human-evaluated severity, with unhelpful critiques being least severe. For topic-based summarization, we again use humans for ground truth, and additionally break ties between helpful critiques using severity.

For synthetic tasks, we again use binary helpfulness ground truth as if it were acting as the human being assisted. This simulates a human judgement in a setting where the humans can easily evaluate critiques, but not the base task answers.

A downside of this approach is that it makes it possible for the model to find flaws by "trying out" critiques it has low confidence in. This means that C effectively gets to leverage a bit of search against ground truth, which D does not get to do. Therefore one could expect a negative CD gap. However this "search effect" should be quite small in the limit of scale with larger models – humans are doing a relatively simple task of critique validation, compared to the difficult task of finding a critique. If the effect exists, it is a genuine advantage of having a human in the loop. For more discussion, see Appendix C.3.1.

5.3 Results

We report our main topic-based summarization gap results in Figure 8 and our synthetic task gap results in Figure 9. For topic-based summarization, we use a random non-zero temperature for the answer, and use the same temperature for critiques. For synthetic tasks, we use temperature 1 for the answer, and temperature 0 for critiques[6].

We find that:

1. We consistently observe positive GD and GC gaps, as expected, except when models are completely unable to do the task.
2. We observe a positive CD gap for topic-based summarization and 3-SAT and negative gap for Addition and RACE.
3. For topic-based summarization, the CD gap is approximately constant across model scale.
4. For most synthetic tasks, CD gap may be decreasing with model size, but the opposite is true for RACE, where critiquing is close to oracle performance (and is easy relative to knowing when to critique).

Overall, this suggests that gaps are task-specific, and it is not apparent whether we can close the CD gap in general.

We believe the CD gap will generally be harder to close for difficult and realistic tasks. For example, on topic-based summarization the discriminator may be able to identify the labeler who gave the answer based on their writing style, and guess that some labelers give more critiqueable answers, without knowing the critiques. This does not happen with synthetic tasks.

We believe much more work on gaps is possible, and present a few more variants and results in Appendix C.3. Overall, we are excited for future study of gaps as a way to measure headroom for pushing critique performance, and as a way to improve methods for scalable oversight.

6 Related work

Scalable alignment experiments. [CSA18] implement iterative amplification for algorithmic tasks. [ICA18] introduce debate and implement a toy version with sparse MNIST classification. [SRE+20, BCOI20, BCS+20, PTP+22] conduct debate-like experiments on realistic tasks (checking claims about movie reviews, physics problems, and reading comprehension), with humans serving as debaters, generally with mixed results. Conversely, [AZWG21] study variants of debate with learned models serving as judges on toy tasks. [WOZ+21] implements a variant of recursive reward modeling [LKE+18] on summarization tasks.

Human assistance with natural language. [LSSC22] use assistance to help humans create demonstrations to create challenging NLI datasets. [ZNC+22] and [PHS+22] use model assistance to find adversarial examples for language model classifications and generations, respectively. [PKF+19] help humans perform passage-based question-answering, without reading much of the passages.

For helping humans with evaluations, [FPP+20] help humans fact-check claims faster and more accurately with natural language briefs. [GSR19] use language models to help humans discriminate whether text was generated by a model.

Critique datasets and models. [TVCM18] introduce a dataset of factual claims, along with supporting and refuting evidence. [KAD+18] introduce a dataset of critical peer reviews. [BCV16] mines disagreements from Twitter, and [ZCP17, PBSM+21] from Reddit. [MST+21] introduce a dataset of story critiques.

For model generated critiques, IBM's Project Debater [SBA+21] trains models to engage in free text debates, including the ability to rebut arguments. Unlike our work, they focus on debating against humans rather than models.

Natural language refinements. Human natural language feedback has been used to improve models in many domains, such as computer vision [RLN+18], program synthesis [EHA20, AON+21], and summarization [SCC+22]. [PTA+21] use large language models to fix security vulnerabilities

[6] We initially tried other settings which did not qualitatively change results but made win rates closer to 50% and error bars larger.

in code. More recently, [WWS+22b] propose using language models' own outputs to improve their answers on math word problems.

7 Discussion

We view our results as a proof of concept for feedback assistance as a solution to the problem of scalable oversight: Even though topic-based summarization isn't actually a hard task for human labelers, in our experiments we still see significant gains from AI assistance in the form of critiques.

7.1 Implications for alignment research

1. **Large language models are already capable enough to meaningfully assist human evaluation** and the scaling trend in Figure 4 suggests that larger models may improve at assisting in evaluating their own outputs. The publicly available InstructGPT models are capable of critiquing well few-shot and even zero-shot (Figure 3). Overall, we believe there is potential to do empirical experiments for scalable oversight with today's models, using schemes similar to reward modeling [LKE+18] or debate [IA19].
2. **Generator-discriminator-critique gaps are promising ways to measure alignment properties of models.** Studying gaps give us insight into quality of base task training signal without training those models (see Appendix C.3). Increasing the critique performance relative to generator and discriminator performance is an under-explored research area, where results should directly translate into better-aligned models. Studying gaps can also happen on smaller models in synthetic domains, like those in Table 3.
3. **Learning from natural language feedback is feasible now.** Feedback in preference learning [CLB+17] is very information-sparse, and humans typically spend several minutes on a comparison yielding a single bit of information. Ideally, models could use human natural language feedback to improve their own outputs [SCC+22]. In Section 4.3, we showed models can now condition on critiques as a form of feedback to improve their own outputs, results corroborated by recent works on "chain of thought" [WWS+22b]. This suggests teaching models with natural language feedback from humans is a very promising direction.

7.2 Limitations

1. **Lack of ground truth.** Our base task of topic-based summarization does not have a robust or objective process for validating the quality of the answers or critiques.
   (a) Labelers may be misevaluating answers, by trusting the model summaries too much or by simply making mistakes.
   (b) Some critiques found by the labelers using assistance were fairly unimportant or nitpicky. Agreement rate on comparisons of critiques (i.e. helpfulness rankings) were no higher than answer comparisons; both were around 75%.
   (c) Misleading critiques of good outputs may be indistinguishable from good critiques of poor outputs.
   (d) More broadly, we do not address how to measure ground truth, which makes this research difficult. Our work relies on labelers, who already make mistakes and will be increasingly unreliable for harder tasks.
2. **Assuming articulable reasoning.** Our overall research direction does not address how to surface problematic outputs where a model cannot put into words what the problem is, which may be a core difficulty of the alignment problem [CCX21]. The CD gap could remain large after much effort using assistance.
3. **Assuming reconcilable preferences.** Critiques as a training signal may not make sense for more inherently subjective tasks, where labelers have differing preferences. It may be impossible to have uncritiqueable outputs (at least without specifying how to resolve disagreements). On the other hand, for subjective tasks having a strong critique model can make it easier to adapt a model to each labeler's individual preferences because it lets them rank the critiques they care about without having to find all of them.