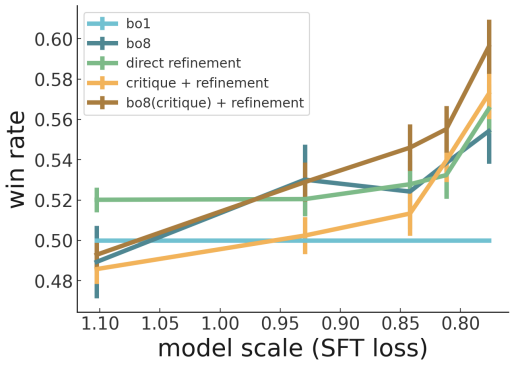
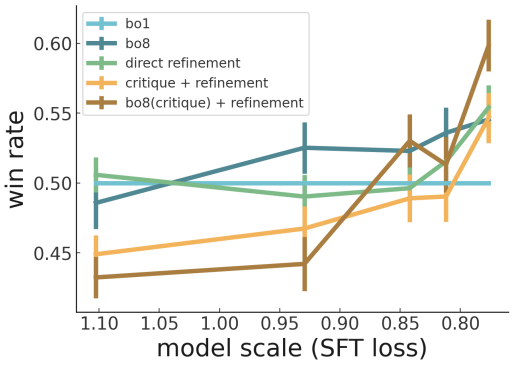
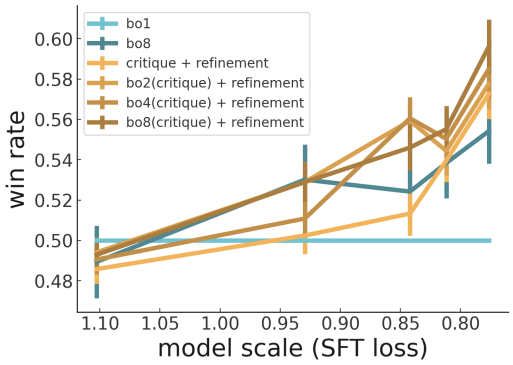


(a) Comparison of critique-conditional refinements to three baselines: the original sample, a direct refinement, and a best-of-8. Small models are poor at refining. For large models, critique-conditional refinements outperform baselines.

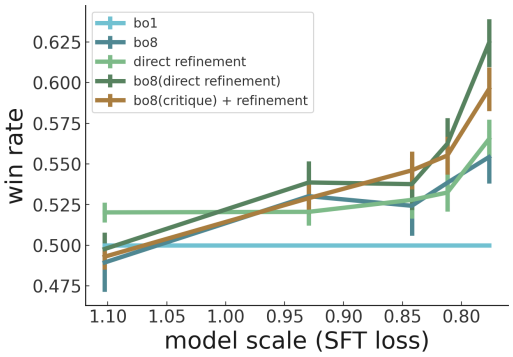


(b) Using "forced" refinements, we see that small models are exceptionally bad at conditional refinements. In this setting, the model has no ability to opt out of critiquing or direct-refining.

Figure 6: Critiques help with refining answers. They are also competitive with direct refinements, and a best-of-8 baseline. However, these are only true at scale. Win rate is measured relative to the original (best-of-1) answer from the same model. All critiques and refinements are generated from the same model as the answer, and all generations are at T=0.5.



(a) Win rate of critique-conditional refinement against the original answer. Better critiques (found via best-of-N against the helpfulness model with increasing N) seem to improve refinements, though results are noisy.



(b) Best-of-8 with direct refinements offers a more competitive baseline that possibly outperforms critique refinements. All 8 refinements are of the same original answer.

Figure 7: Critique refinement and direct refinement scaling with rejection sampling. Figure 7a assesses conditional refinements optimizing the critique against helpfulness score, whereas Figure 7b assesses direct refinements optimizing the refinement against critiqueability score. Win rate is measured relative to the original (best-of-1) answer from the same model. All critiques and refinements are generated from the same model as the answer, and all generations are at T=0.5.

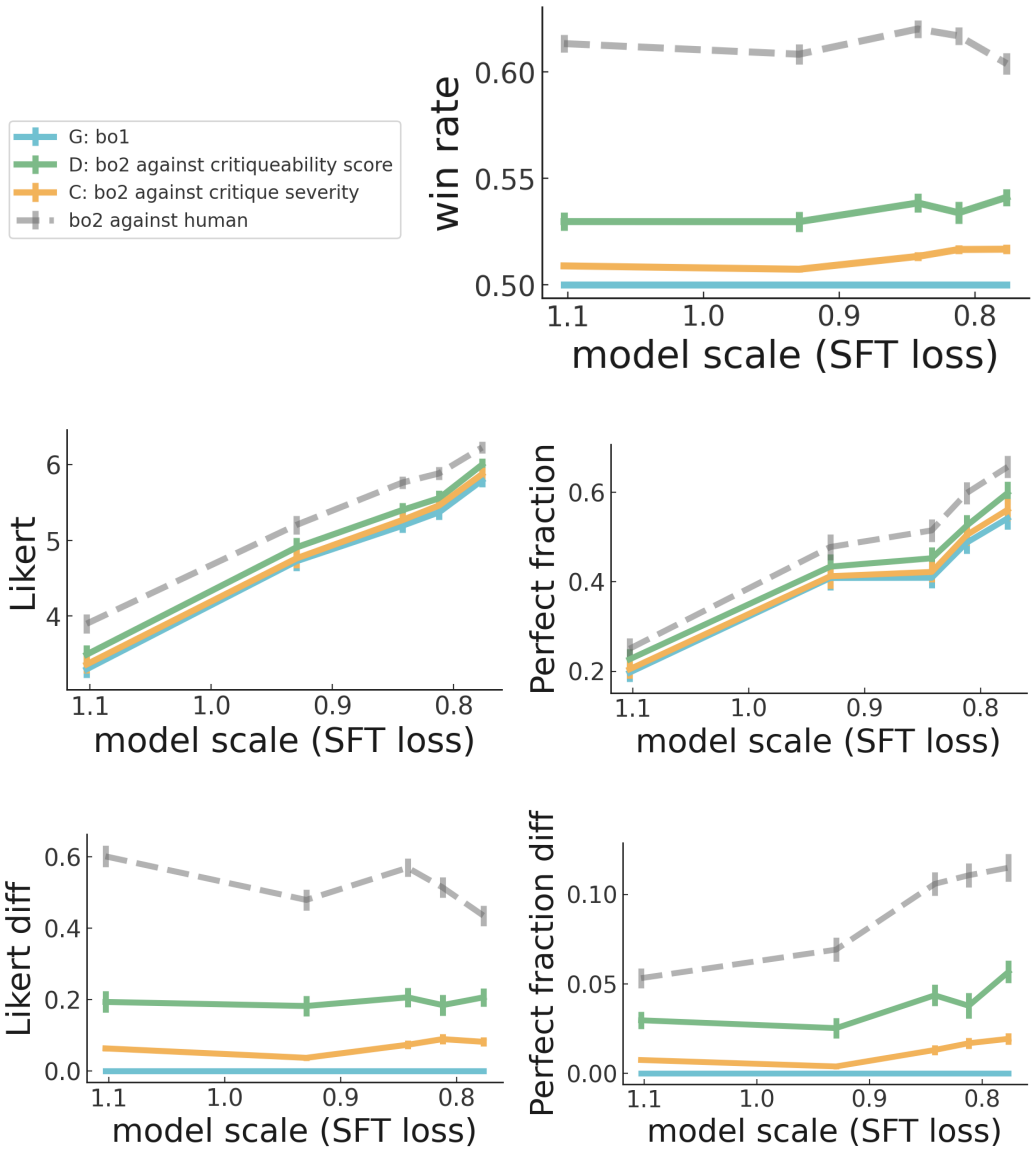


Figure 8: GDC gaps for topic-based summarization, using humans as ground truth. We measure sample quality using various metrics. "Diff" metrics subtract out the values for the generator. Note that best-of-2 against human win rate against best-of-1 would be exactly 75% if not for labelers marking ties. Overall, GD and GC gaps may be slightly increasing, but CD gap is positive and shows no trend.

In this section, we present one such way of measuring and our results using it.

5.2 Measuring gaps

We propose comparing these tasks to each other using the following methodology:

- 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?

For measuring C, we essentially use critiques as a discriminator: to judge an answer we generate a critique and consider the answer poor if any critique is valid and severe, according to a human.

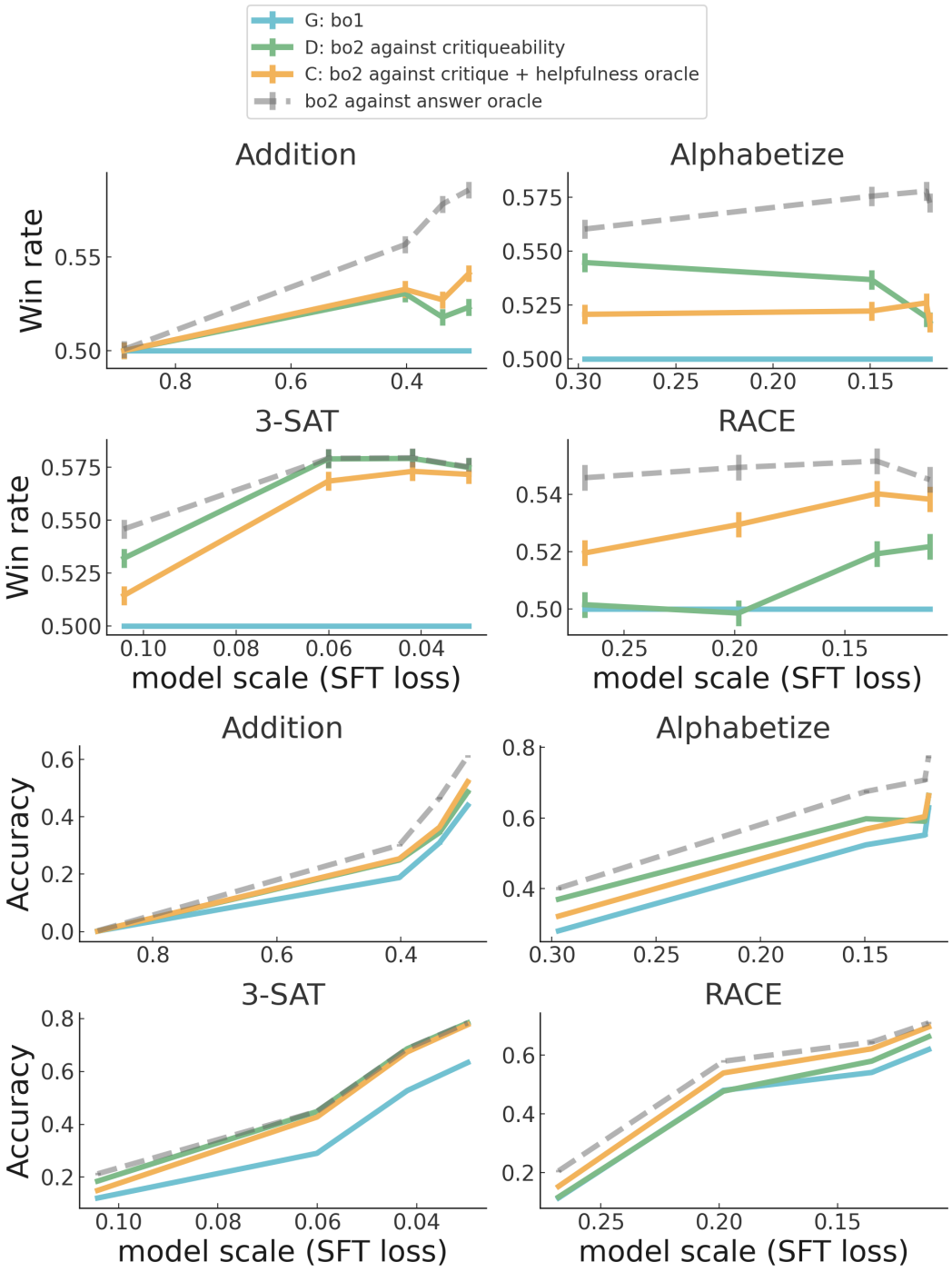


Figure 9: GDC gaps for synthetic tasks, using an oracle as ground truth. We also show the oracle best-of-2 discriminator. Note that for binary tasks, win rate is a linear transformation of accuracy gaps. We do not see consistent trends with CD gaps.